Evaluating Parallel Corpora

Assessing Utility for Use with Translation Memory Systems in Government Settings

The Tenth Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2012)

Authors (in alphabetical order):

Sergey Blok
Michael Bloodgood
Petra Bradley
Ryan Corbett
Michael Maxwell

Erica Michael (presenter)
Peter Osthus
Paul Rodrigues
Benjamin Strauss

UNIVERSITY OF MARYLAND
CASL
CENTER FOR ADVANCED STUDY OF LANGUAGE

29 October 2012

CASL Overview

- What is CASL?
  - The only University Affiliated Research Center devoted to the study of language
- What do we do?
  - Conduct independent, empirically based science guided by the strategic needs of the USG
- Who are we?
  - 190+ researchers in fields such as linguistics, cognitive science, computer science, and second language acquisition

UNIVERSITY OF MARYLAND
CASL
CENTER FOR ADVANCED STUDY OF LANGUAGE

Goal of the Project

- Per request from the National Virtual Translation Center (NVTC), assess goodness-of-fit of parallel corpora with customer's material to be translated (MTBT)
  - Based solely on contents of the corpora
  - Based on how the corpora will be used with various Translation Memory (TM) systems and other types of translation technology (e.g., machine translation, terminology management)

Objectives and Impact

- Develop evaluation heuristic to:
  - Identify the most suitable parallel corpora
  - Identify the type of additional corpora needed
- With increasing availability of resources, translators need good methods of identifying and building corpora and other resources that are most suitable to the task at hand.

Parallel Corpora

- Based on pairs of translated documents
  - Aligned by segments to ensure equivalence of meaning across pairs
- Can be used to create Translation Memory (TM) vaults in standard TMX file format
  - System presents potential matches from vault for new translation task
  - User can select degree of match (0-100%)
  - User can access one or more vaults at once

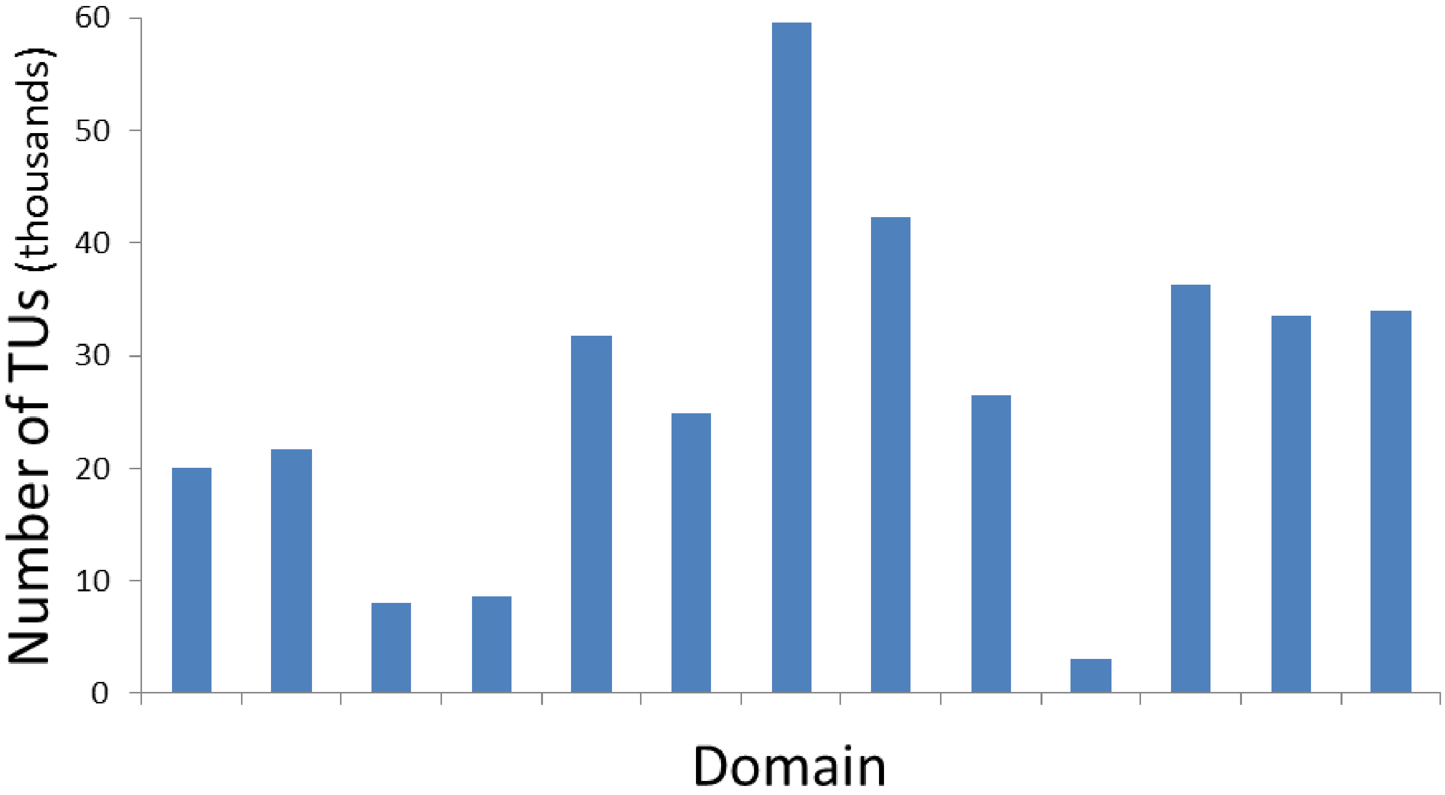
Key Features of TM Vaults

- **Size**
- Segmentation
  - Size of translation unit (TU)
- Goodness-of-fit
  - Language (including dialect, location, etc.)
  - Genre (e.g., journal article, letter, speech)
  - **Domain (e.g., legal, medical, technical)**

TM Curator Project Overview

- NVTC creating TMX files from existing parallel corpora to be used for customer’s translation tasks
  - Aligning pairs of journal articles that have already been translated Chinese → English
  - Tagging each article with one or more of 13 domains of interest

TM Curator Project Vaults



NVTC’s Research Questions

- How do you decide whether a given vault will be useful for a new translation task?
  - How “similar” is the vault to the new material?
  - How much coverage does the vault provide?
- What are the best ways to improve the usefulness of a vault?
  - Is there a tradeoff between vault size and domain specificity?

Research Plan (in progress)

- Identify metrics for assessing similarity and coverage
  - Scores should differentiate between similar and dissimilar sets of documents; i.e., be better within a domain than across domains.
  - Cross-domain scores may vary with the similarity/distinctiveness of the domains.
    - May provide information about which vaults could be combined

Test Conditions

- MTBT always 1,000 segments
- Tested five different vault sizes
  - 1,000 / 5,000 / 10,000 / 20,000 / 30,000 segments
  - Not all domains tested at all sizes
- Intra-domain comparisons
- Cross-domain comparisons

Outcomes of Interest

- Number of matches between MTBT and TM vaults
- Quality of matches
- Usefulness of matches from the perspective of a human translator
- Ultimately …
  speed and quality of translation

Metrics (1)

- Computed at the level of the segment
  - Percent match
  - Weighted percent match
  - Longest common substring
  - Edit distance
- Computed at the level of the corpus
  - Coverage

Metrics (2)

- All can include pre-processing with stop word filtering
- All were computed using our own algorithms; many TM systems compute similar metrics

Percent Match

- Percent of tokens in the MTBT segment that are also found in the vault segment

MTBT Segment	Vault Segment
The man walked to work	After work the woman walked her dog
3 of 5 tokens in the MTBT segment are also in the vault segment, so Percent Match = 3/5 = 60%.	

Aside: Averaging

- For each segment in the MTBT, the algorithm searches the entire vault for the segment that provides the highest percent match.
- Each data point represents the average of the best matches across all 1,000 segments in the MTBT corpus.

Weighted Percent Match

- Weights the words by their inverse document frequency (IDF), then computes the percent match using the IDF weights

$$Weighted\ Percent\ Match = {\sum_{i \in T \cap V} idf_i} \Big/ {\sum_{i \in T} idf_i}$$

(T is the tokens in the MTBT segment; V is the tokens in the vault)

- Benefit: gives increased weight to words that occur less frequently

Longest Common Substring

- Longest sequence of words common to both the MTBT and vault segments, divided by length of the MTBT segment

<table>
<tr><th>MTBT Segment</th><th>Vault Segment</th></tr>
<tr><td>Yesterday the man with the red car drove to work</td><td>A man I met yesterday drove to work in a red car</td></tr>
<tr><td colspan="2">Longest common substring is 3 words and MTBT segment is 10 words, so score is 3/10 = 30%.</td></tr>
</table>

Edit Distance

- Number of insertions, deletions, and substitutions made to the MTBT segment to transform it to the vault segment
  - Formula includes transformation such that fewer edits = higher score
    - Maximum score of 1 when MTBT segment is identical to vault segment; artificial lower bound of 0 so score is never worse than if calculated using a blank vault segment
    - Presented as percentage (original score x 100)

Coverage (1)

- Percentage of terms (i.e., unique tokens) in the MTBT that appear in any segment in the vault
  - NOT averaged across segments; computed by comparing an entire MTBT corpus to an entire vault

$$Coverage\ Score = 100 * \frac{|Terms\ in\ MTBT \cap Terms\ in\ Vault|}{|Terms\ in\ MTBT|}$$

Coverage (2)

MTBT Segments	Vault Segments
The dog is fast ① ② ③ ④ The house is red ⑤ ⑥	The car is red ① ③ ⑥ The dog is running ②
The MTBT contains 6 unique tokens, 4 of which are also in the vault, so the Coverage Score = 100 x 4/6 = 66.6.	

Stop Word Filtering

- Used to remove stop words (e.g., articles) before metrics are computed
  - Expected to improve usefulness of match statistics through matching only on content words
  - Can be used before computing any of the metrics described

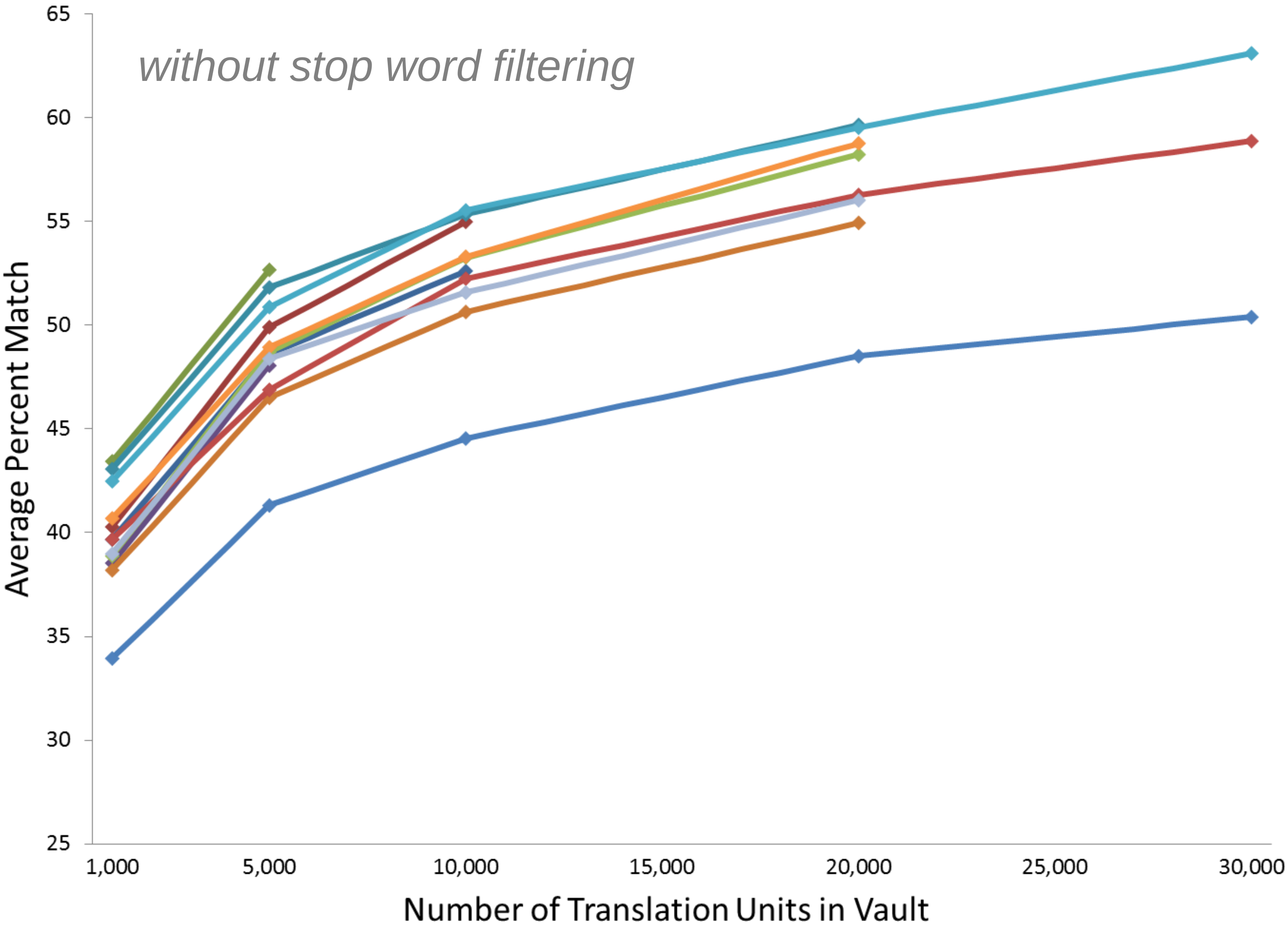
Preliminary Results

- Questions we have begun to address
  - **Size:** When do you hit diminishing returns in adding material to your vault?
  - **Domain:** Which domains serve as the best vaults for a given set of MTBT?

Vault Size Results

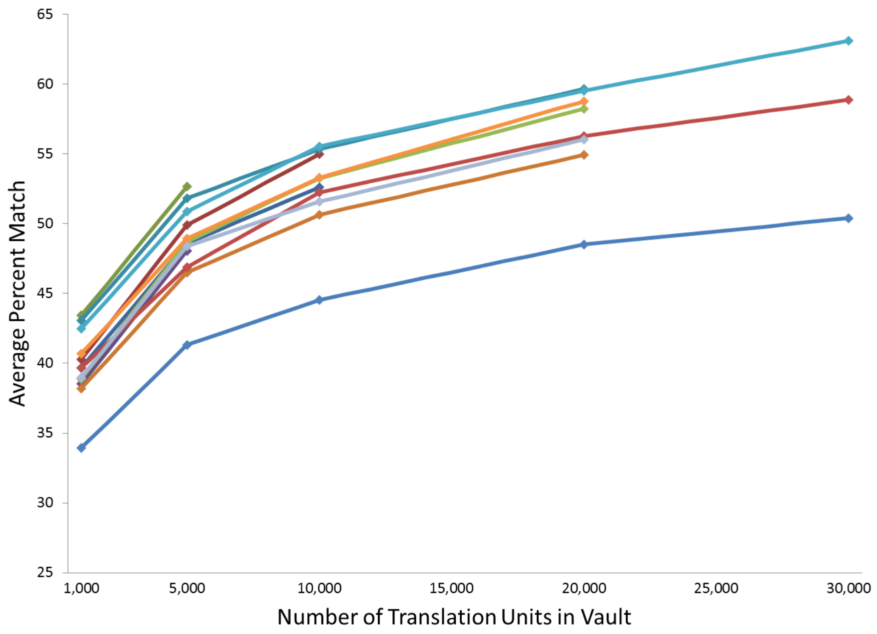
- Intra-domain comparisons
  - For each domain, MTBT (1,000 segments) compared to vaults of varying size from the same domain
  - Reminder: For most metrics, each data point represents the average of the best matches across all 1,000 segments in the MTBT corpus.

Effect of Vault Size on Percent Match (1)

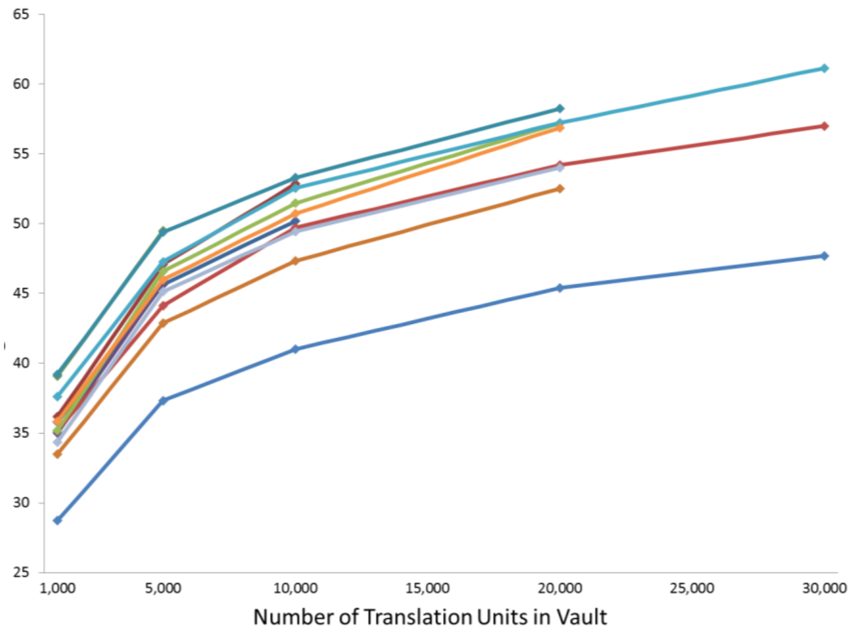


Effect of Vault Size on Percent Match (2)

without stop word filtering

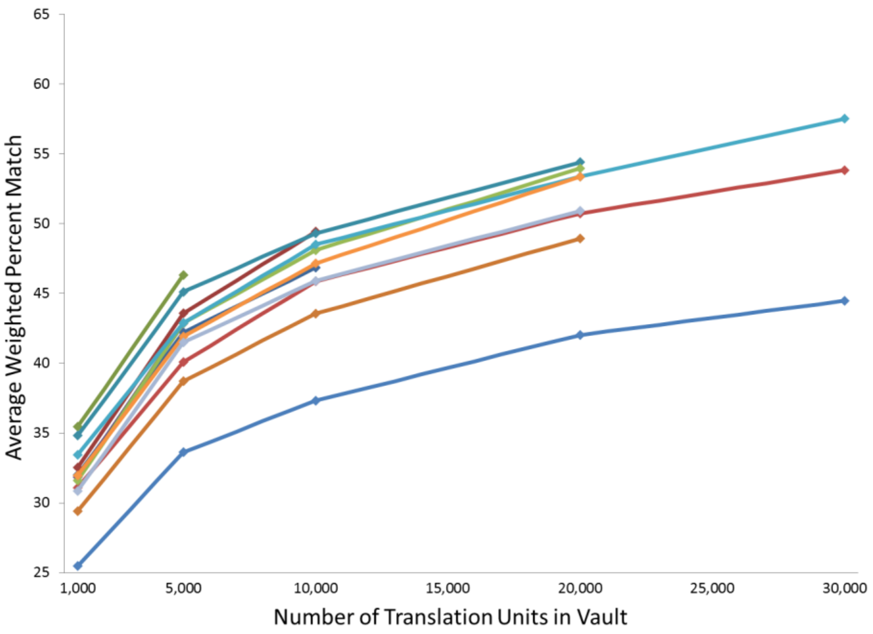


with stop word filtering

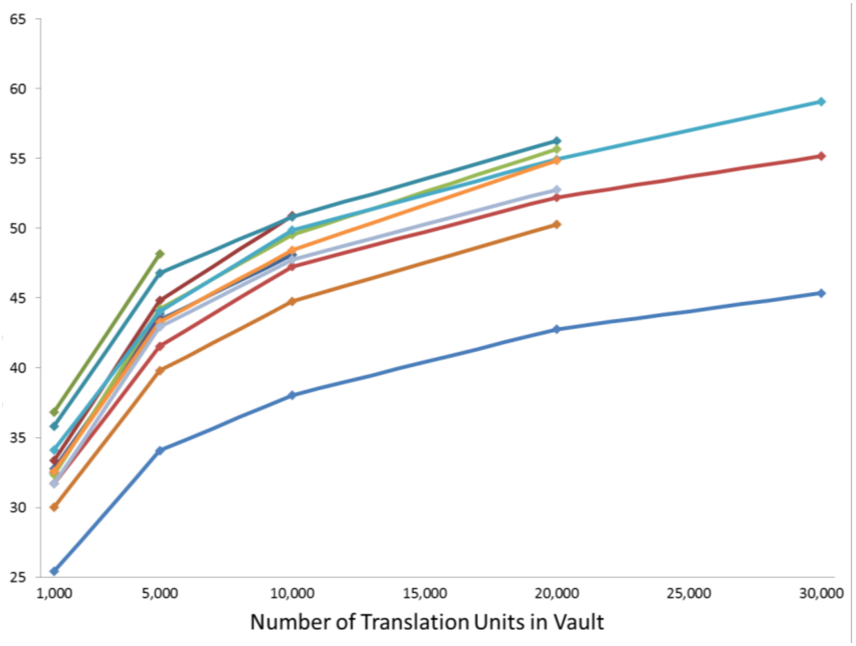


Effect of Vault Size on Weighted Percent Match

without stop word filtering

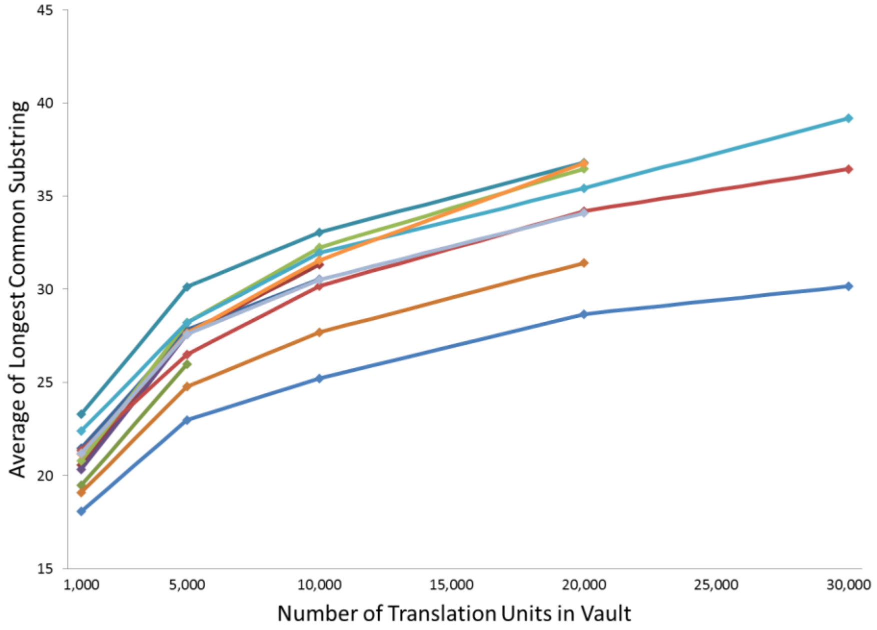


with stop word filtering

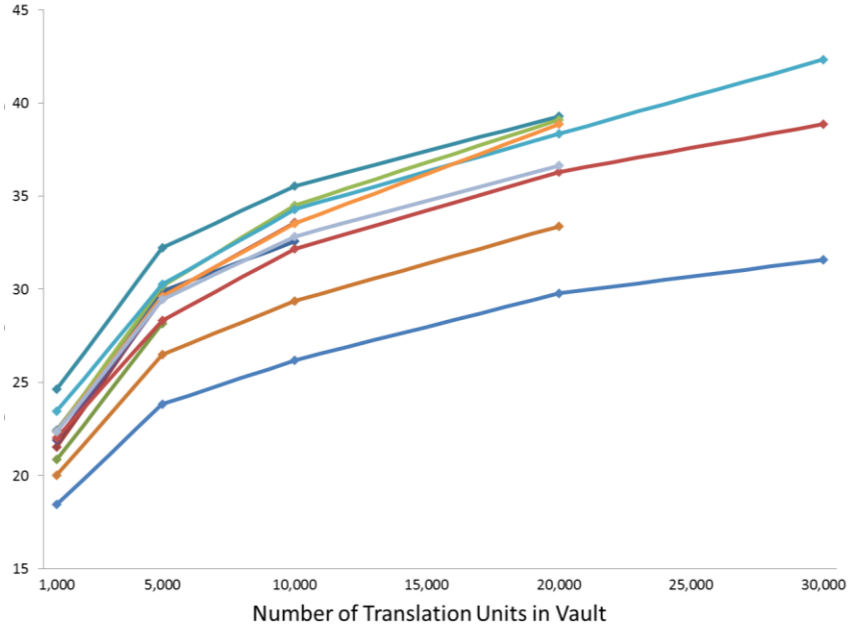


Effect of Vault Size on Longest Common Substring

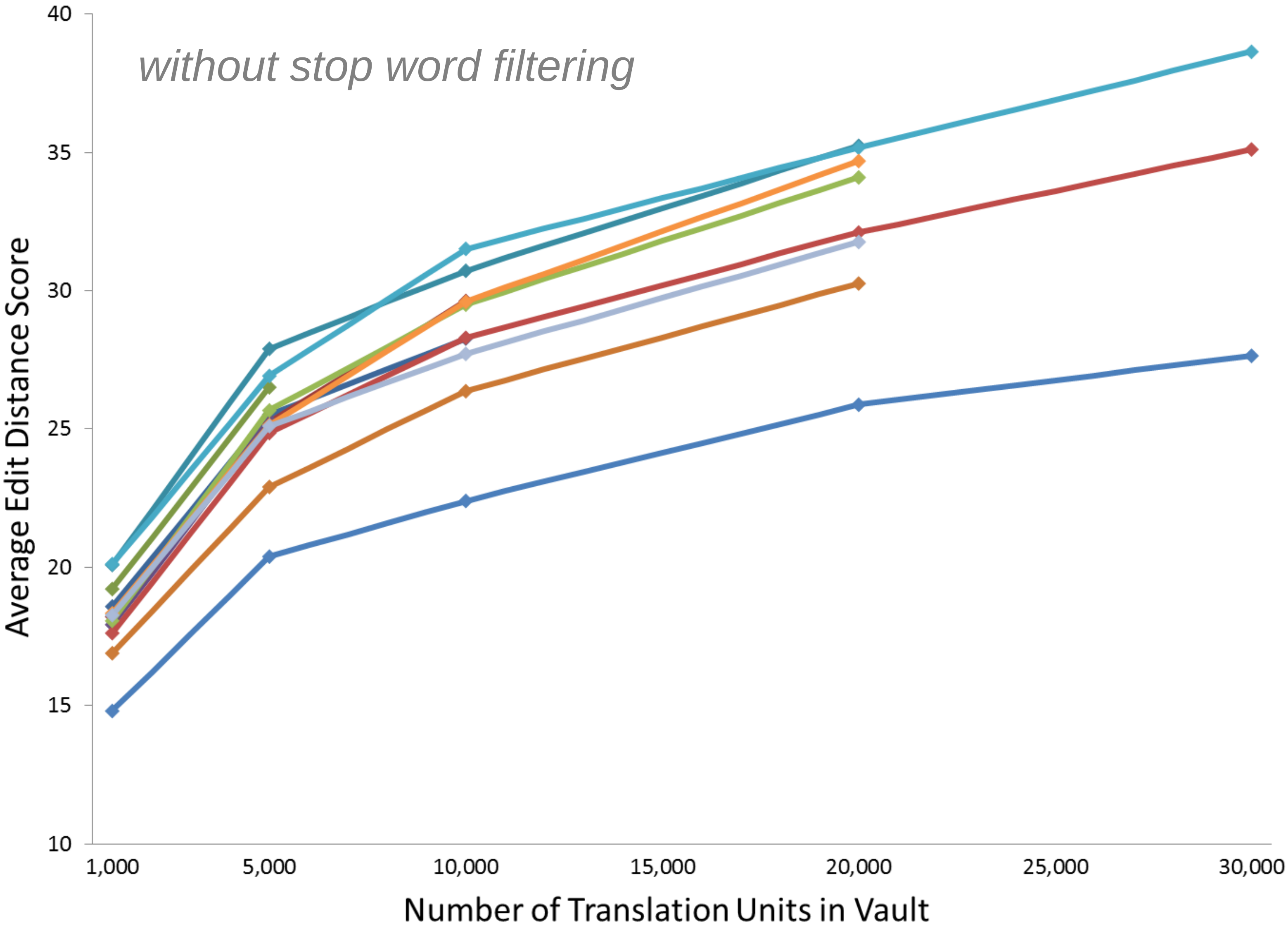
without stop word filtering



with stop word filtering

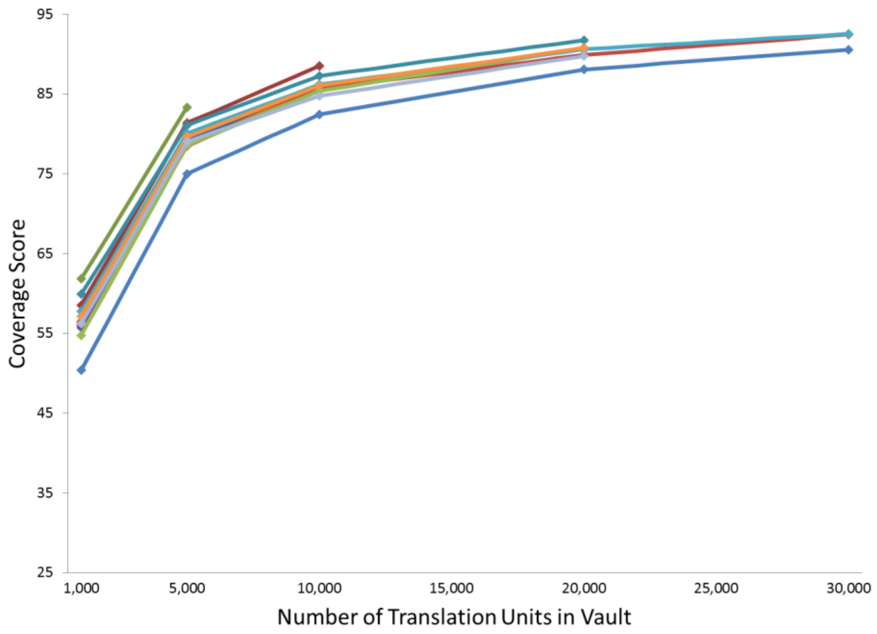


Effect of Vault Size on Edit Distance

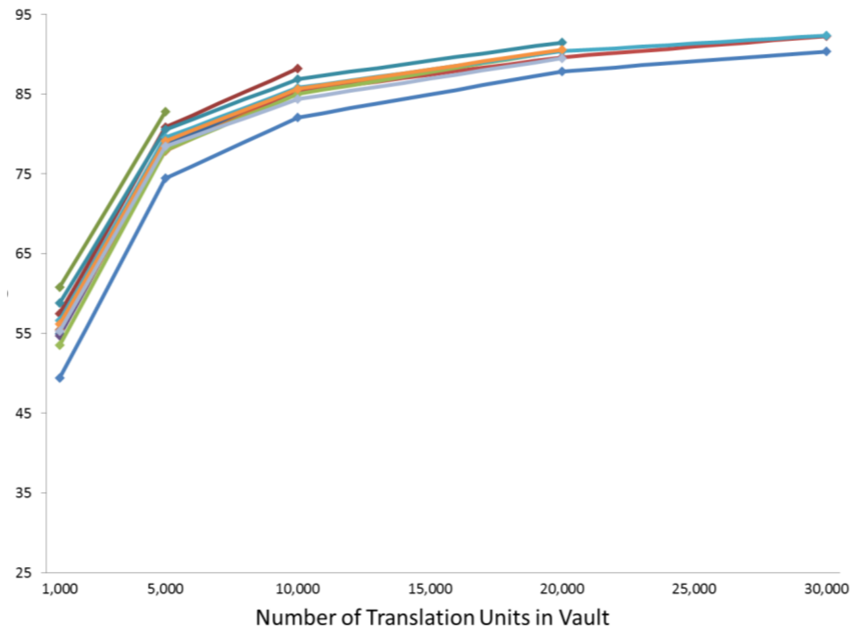


Effect of Vault Size on Coverage

without stop word filtering



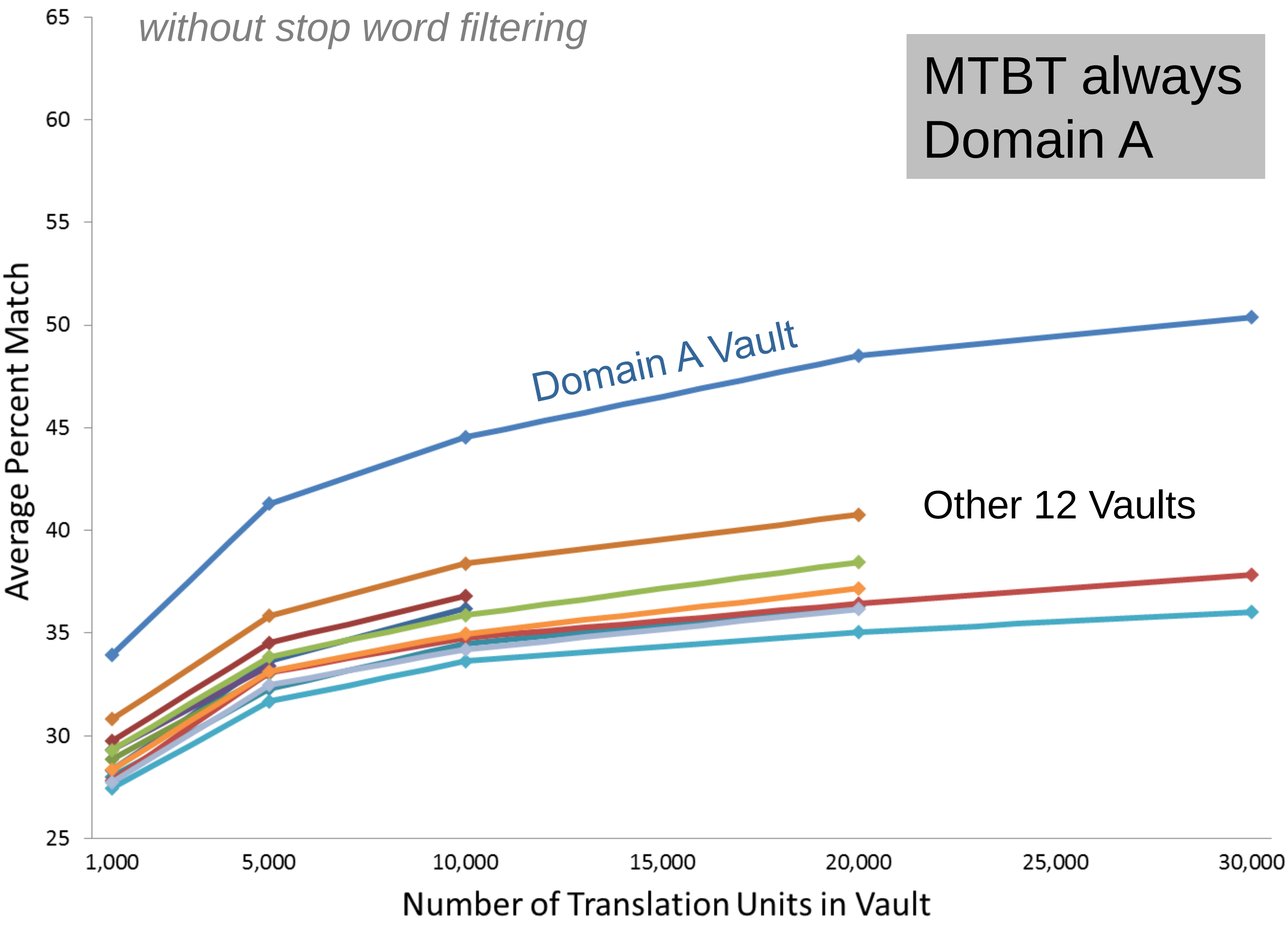
with stop word filtering



Domain Specificity Results

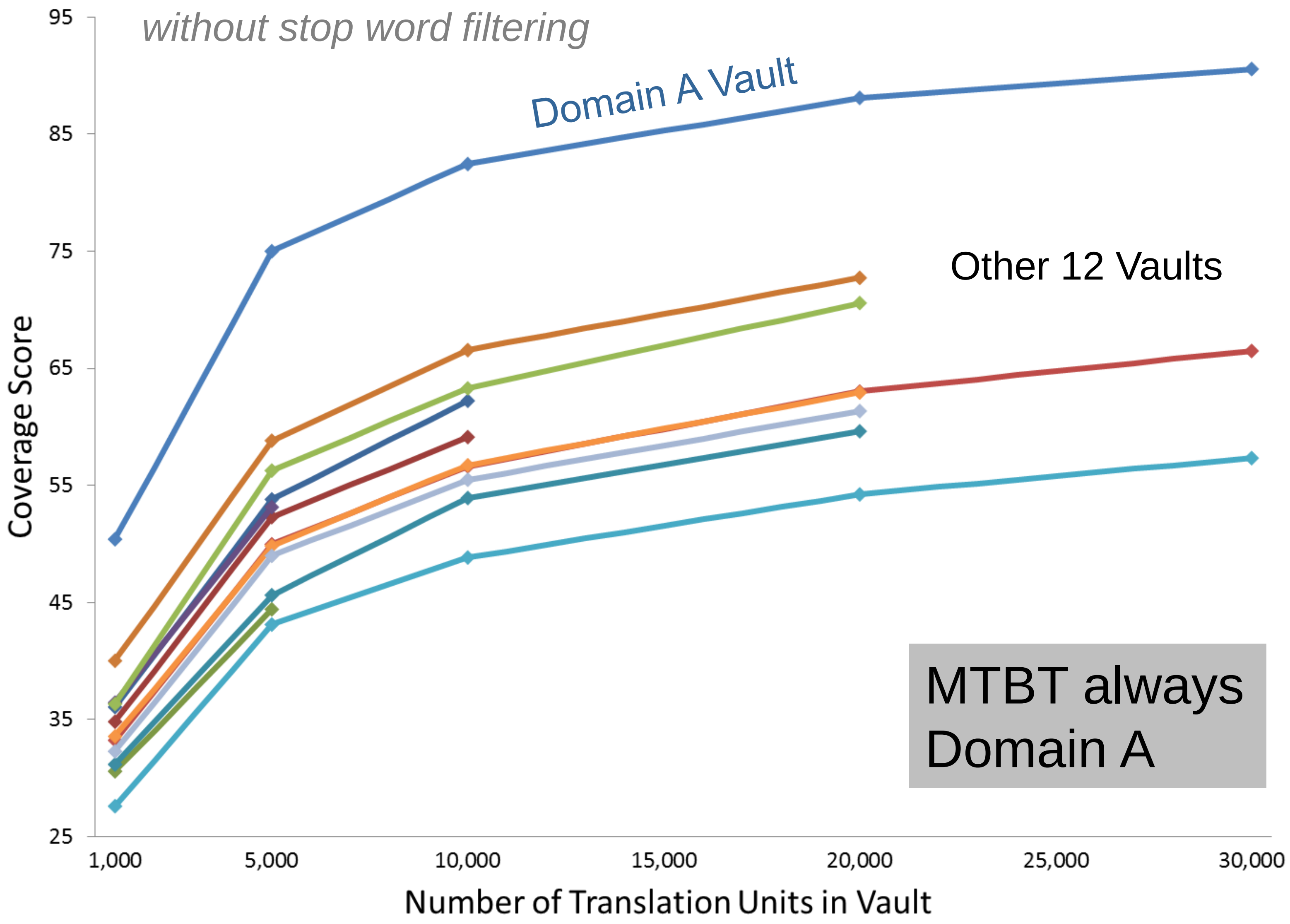
- Cross-domain comparisons
  - For each domain, MTBT (1,000 segments) compared to vaults of varying size from all 13 domains
  - For current examples, MTBT always from Domain A

Percent Match Across Domains



MTBT always Domain A

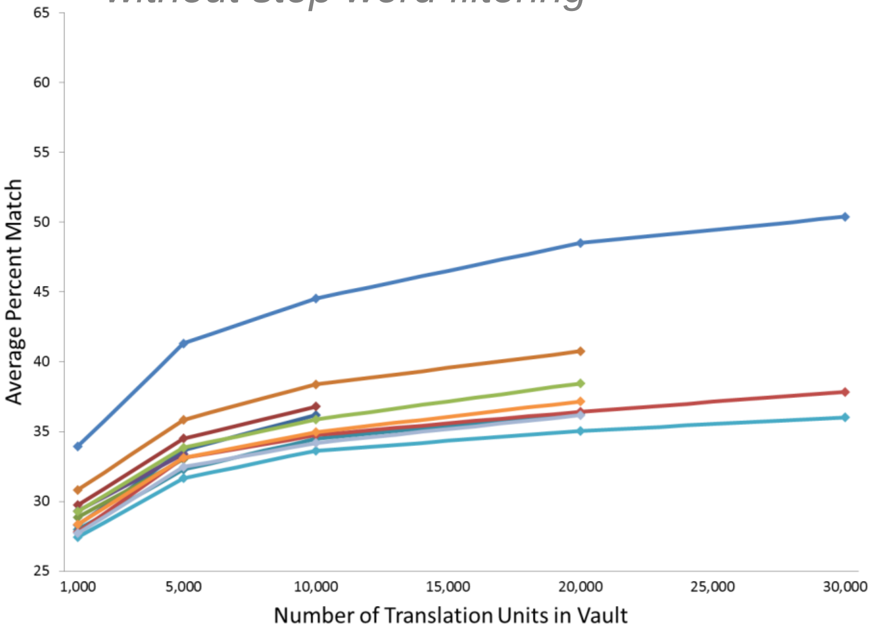
Coverage Across Domains



Percent Match and Coverage Across Domains

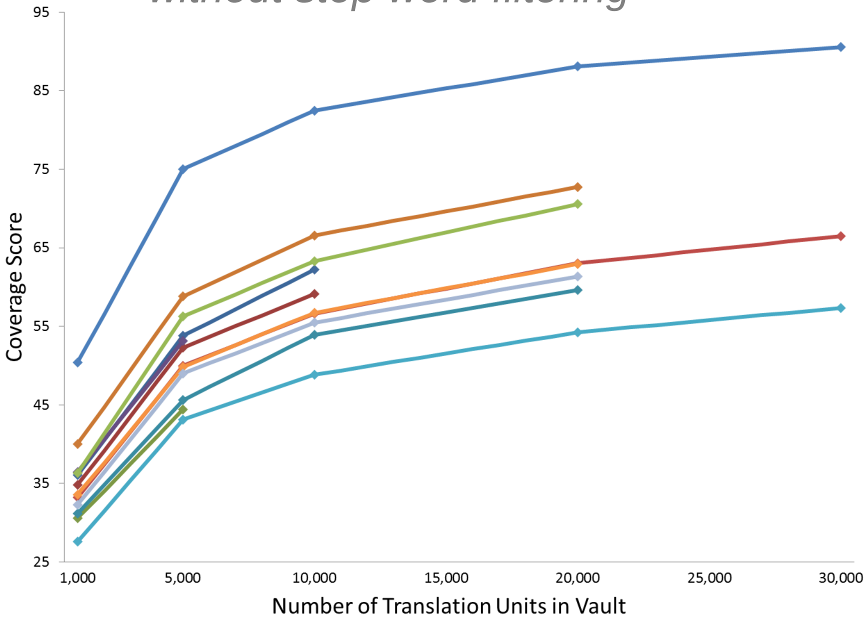
Percent Match

without stop word filtering



Coverage

without stop word filtering



MTBT always Domain A

Domain Similarity

Domain	A	B	C	D	E	F	G	H	I	J	K	L
A	1.0	0.047	-0.053	-0.063	-0.019	-0.045	-0.013	-0.054	-0.032	-0.010	-0.046	-0.102
B		1.0	-0.076	-0.054	0.033	0.016	-0.094	-0.028	-0.058	-0.084	**0.129**	-0.139
C			1.0	-0.080	**0.437**	-0.053	-0.110	-0.077	-0.002	-0.096	-0.089	-0.100
D				1.0	-0.065	-0.104	-0.102	-0.081	-0.022	**0.192**	-0.062	-0.190
E					1.0	0.026	-0.130	-0.089	-0.051	-0.122	-0.073	-0.138
F						1.0	-0.106	-0.071	0.010	-0.087	-0.084	-0.093
G							1.0	**0.064**	-0.069	-0.123	-0.145	-0.044
H								1.0	-0.059	-0.110	-0.172	-0.097
I									1.0	-0.073	-0.077	-0.061
J										1.0	-0.090	-0.203
K											1.0	-0.172
L												1.0

Implications

- What do you do if you need to translate something from a domain that doesn’t have a large vault?
  - Use a small vault?
  - Use a vault from a different domain or combination of domains?
    - Beware of polysemy!

Next Steps (1)

- Examine metrics across combinations of domains
- Compare our metrics to those generated by various TM systems
- Explore applications of these metrics to other types of translation technologies beyond TM

Next Steps (2)

- Assess usefulness of metrics for human translators
  - Which metrics are most likely to lead translators to the best vault and/or the best match for a given segment?
- Investigate the use of automatic and/or manual domain tagging to facilitate vault creation and selection

Thank you!

Questions? Comments?

For more information, please contact Erica Michael:

emichael@casl.umd.edu

301-226-8868