

# Comparative Evaluation of Translation Memory (TM) Retrieval of Arabic-English Sub-segment Fragments

---

AsLing: TC 42 2020

19 November

Khaled ben Milad (Alamary), Swansea University

# Retrieval of Translation Memory System

- Translation Memory (TM) retrieval is a process of recalling previous human translation records from a database
- The TM system allows a translator to re-use highly similar segments which are calculated to be of potential use in translating source segments.

# Matching Metrics of Translation Memory Systems

- TM matching measurement is a process of comparing a source segment against TM source.
- Most matching metrics based purely on character-string similarity - Levenshtein edit distance (addition, deletion, substitution operations)
- Matching scores are then calculated according to the measurement of the three-operation edit between the two segments at either a word level or character level.

## Related Studies

Baquero and Mitkov ([2017](#)) performed an investigation in terms of detecting similarities in sentences with minor revisions.

- Transformation some linguistic rules (English-Spanish)
- memoQ, OmegaT., SDL Trados Studio2017, Wordfast

### Results

- TM matching algorithm showed inability to return matching above the default matching threshold (75%).
- High rate of errors for all TM systems reported with syntactic transformations.

# Free Word Order & Semantic Similarity in Arabic

- Free word order in Arabic (semantically identical but structure different)
  - I. **سيفرح** **الطفل** **بلعبته الجديدة** / sayafrah altifl bilaebatih aljadida/
  - II. **الطفل** **سيفرح** **بلعبته الجديدة** / altifl sayafrah bilaebatih aljadida/
  - III. English translation for (I and II ) is: The child will be glad about his new game.
- Question: if one of these sentences were given to a translator as a source text but their TM database contained the other version, would the TM metrics return segments including a sub-segment fragment in high scores?

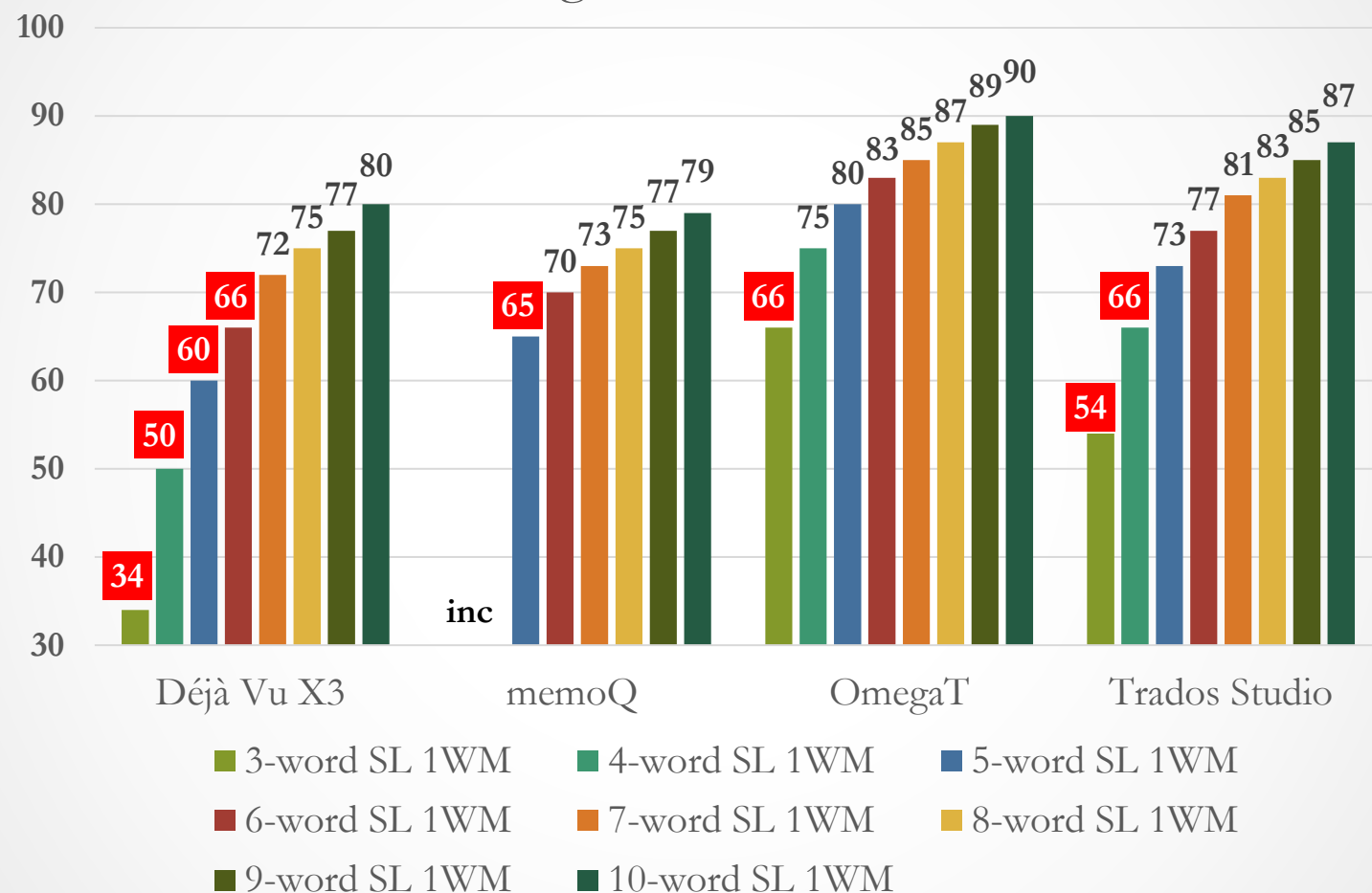
# Experimental Setup

- Building a test set: 95 extracting segments from the open Arabic-English resource MeedanMemory.
- Sentence Length: 3 to 10 words. / String Move: a one-word event unit(1WMU), 2WMU, 3WMU, and 4WMU./ 3 samples used in each event.
- The file to be translated is uploaded to the CAT systems, while the MeedanMemory extract is imported to the TM as a TMX file.
- CAT Tools used: Déjà Vu X3, memoQ 9.5, Memsource Cloud, OmegaT, SDL Trados Studio 2019.
- Setting the minimum match threshold at (70%).

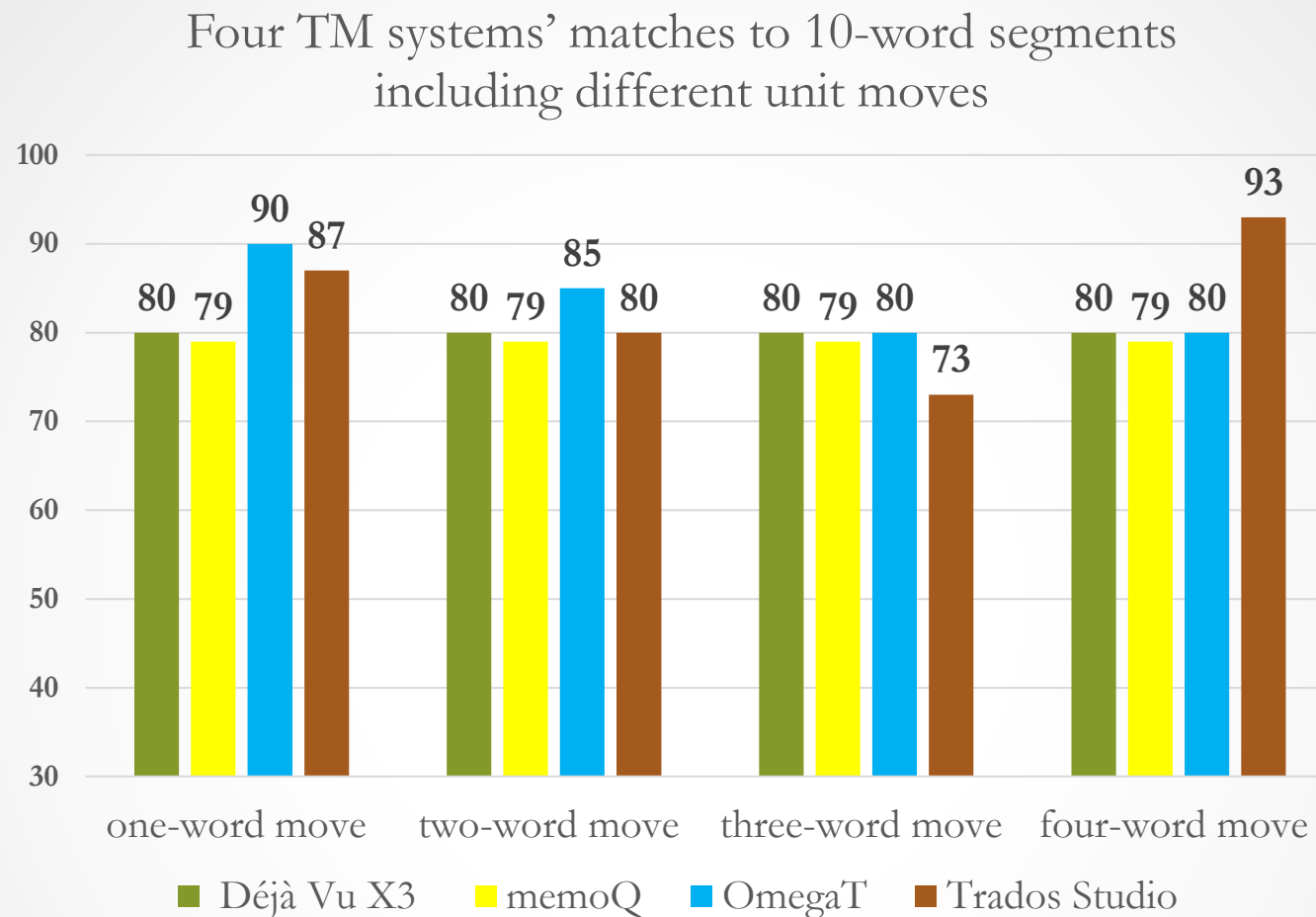
## Findings 1: Consistent Scores.

Matching scores  
reduced as the length  
of the segments  
decreased.

Four TM systems' matches to segments  
including a one-word move



**Findings 2:**  
The string move is treated either as a number of discrete words, or as an undifferentiated block.



## **Memsources Cloud: Inconsistent scores**

Retrieval values are classified into 2 groups:

- I.** The retrieval of segments of 14-30 characters scored a match value below 70% (i.e. Short segments returned in a low match).
- II.** The retrieval of segments of 31-70 characters scored a match value above the threshold (i.e. long segments returned in a high match).

## Conclusion

- TM use the string of surface forms only, no linguistic knowledge.
- Low recall v. high precision
- The current TM matching mechanism has a more negative impact on short sentence routines than on longer ones.

## Proposals & Suggestions

Proposals for developers to improve the matching scores

- Trados Studio deals statistically with a four-word unit move as one chunk, which provides good results.

Suggestions for translators to overcome limitations

- Use a lower match threshold such as 65%.

# References

Baquero A. S. and Mitkov R. 2017. [Translation Memory Systems Have a Long Way to Go.](#)

Bloodgood M. and B. Strauss (2014). [Translation memory retrieval methods.](#) In Proceedings of the 14th Conference of EACL. Gothenburg, Sweden

Gupta, R., Orasan, C., Liu, Q. and R. Mitkov. 2016b. [A Dynamic Programming Approach to Improving Translation Memory Matching and Retrieval using Paraphrases.](#)

<https://atril.com/>

<https://github.com/meedan/news-memory>

<https://omegat.org/>

<https://www.memoq.com/memoq-versions/memoq-9-5>

<https://www.memsource.com/>

<https://www.sdltrados.com/>

Reinke, U. (2013) [tate of the Art in Translation Memory Technology"](#). Translation: Computation, Corpora, Cognition, 3(1).

Simard M, and Fujita A (2012) [A Poor Man's Translation Memory Using Machine Translation Evaluation Metrics.](#) In: Proceedings of AMTA.

Somers, H. (2003) [Translation memory systems.](#) In Somers, H. (ed.), Computers and Translation: A Translator's Guide. Amsterdam/Philadelphia: John Benjamins.

Timonera K and Mitkov R (2015) [Improving Translation Memory Matching through Clause Splitting.](#)