Improve Machine Translation for Cross-lingual Search in E-commerce

With selected Translation Memory using Search Signals

Bryan Zhang

amazon

Content

1. Multilingual Search in E-commerce
2. Machine translation (MT) in cross-lingual search
3. Translation memory (TM)
4. Improve neural machine translation (NMT) with translation memory
5. Optimal translation memory selection workflow
6. Experiment and result analysis

Multilingual Product Search in E-commerce

- **Multilingual search capability** is essential for modern e-commerce product discovery.

- Localization of e-commerce sites have led users to expect search engines to handle **multilingual queries.**

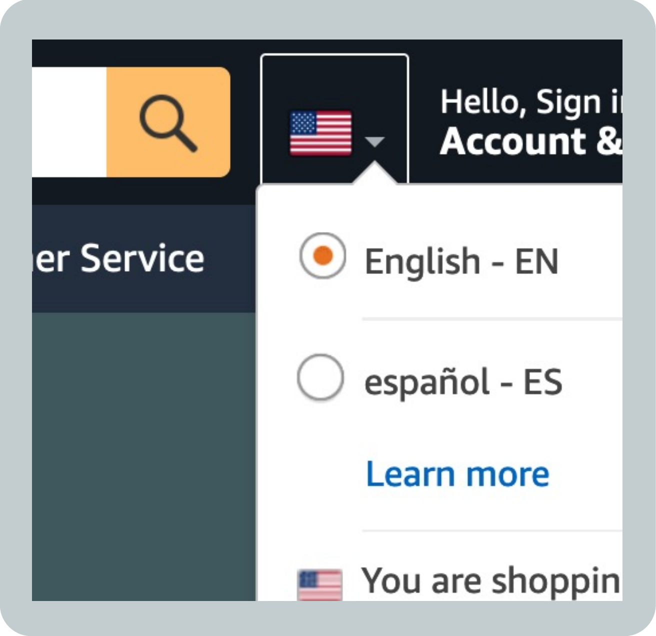
Multilingual Product Content in E-commerce

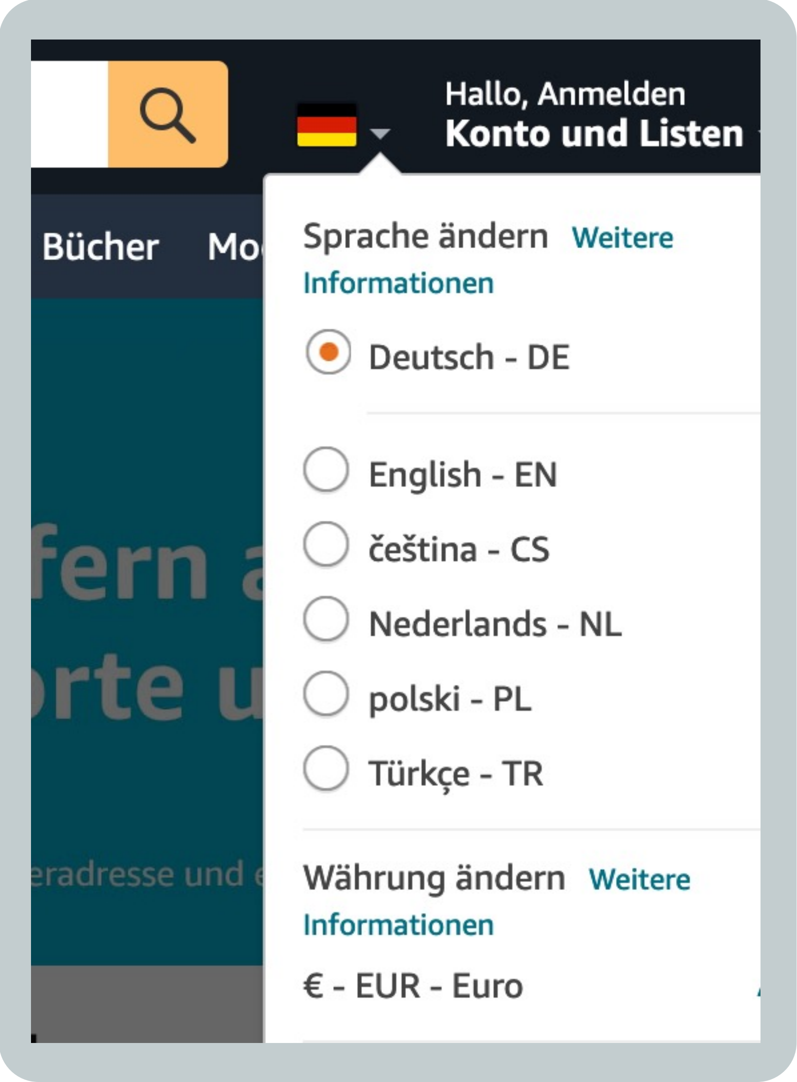
- The classic system that changed gaming history is back!
- Get into spirit for the 35th anniversary of Super Mario Bros. with Game & Watch: Super Mario Bros.
- This special system includes: Super Mario Bros., Super Mario Bros.: The Lost Levels, Ball (Mario version) and a digital clock
- The original Game & Watch system was released in Japan in 1980 and was the very first handheld gaming console created by Nintendo.

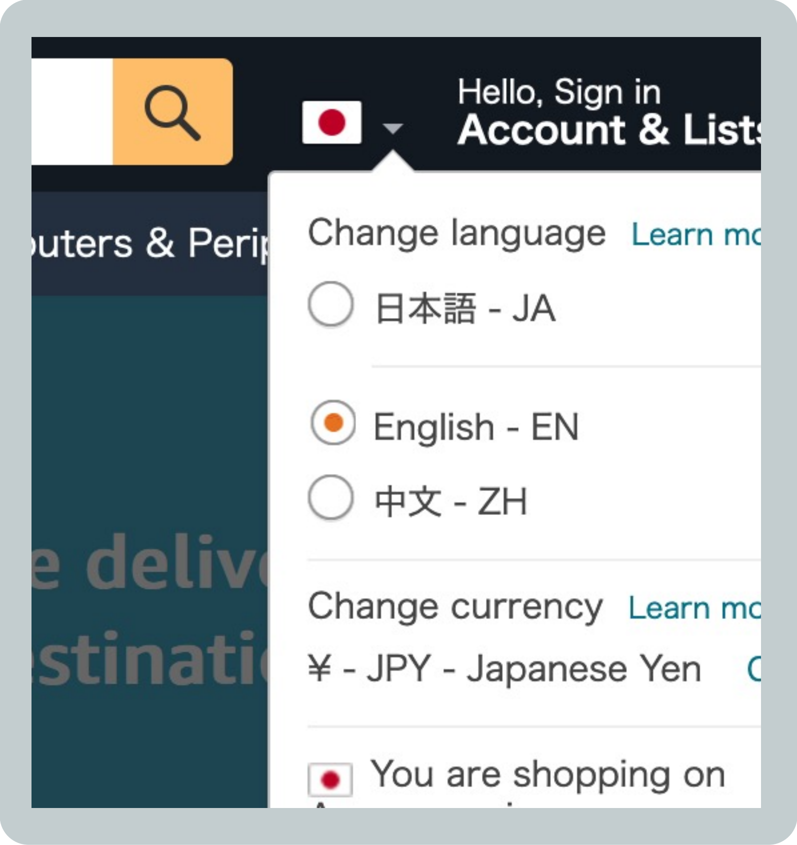
- ¡El sistema clásico que cambió la historia del juego está de vuelta!
- Entra en espíritu para el 35 aniversario de Super Mario Bros. con Game & Watch: Super Mario Bros.
- Este sistema especial incluye: Super Mario Bros., Super Mario Bros.: The Lost Levels, Ball (versión Mario) y un reloj digital
- El sistema original Game & Watch fue lanzado en Japón en 1980 y fue la primera consola de juegos portátil creada por Nintendo.

- Das klassische System, das die Spielgeschichte verändert hat, ist zurück!
- Machen Sie sich zum 35. Jubiläum von Super Mario Bros. mit Game & Watch: Super Mario Bros
- Dieses spezielle System beinhaltet: Super Mario Bros., Super Mario Bros.: The Lost Levels, Ball (Mario Version) und eine digitale Uhr
- Das originale Game & Watch System wurde 1980 in Japan veröffentlicht und war die erste Handheld-Spielkonsole, die von Nintendo geschaffen wu

Multilingual queries in E-commerce







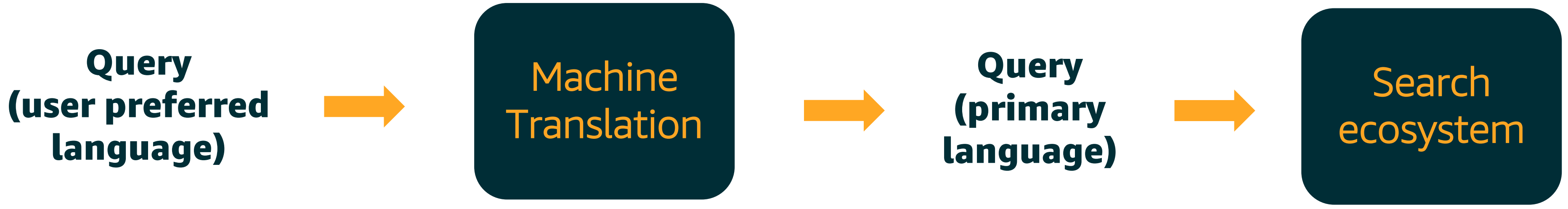
Machine translation (MT) in cross-lingual search

E-commerce search engines typically support multilingual search by cascading a machine translation step before searching the index in its primary language.

Query (user preferred language) → Machine Translation → **Query (primary language)** → Search ecosystem

Machine translation (MT) in cross-lingual search

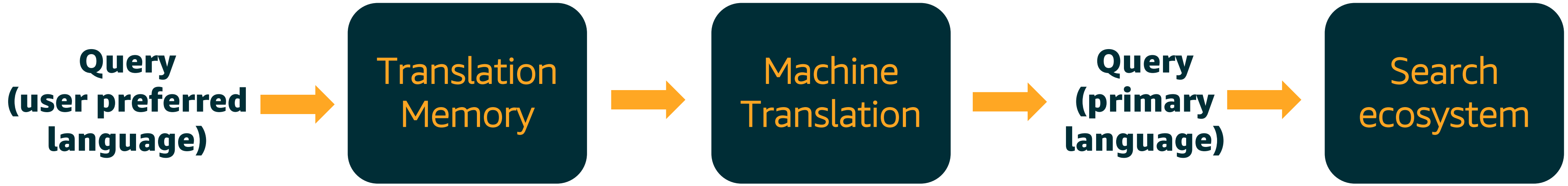
E-commerce search engines typically support multilingual search by cascading a machine translation step before searching the index in its primary language.



In practice, search query translation usually involves a translation memory matching step before machine translation.

Machine translation (MT) in cross-lingual search

E-commerce search engines typically support multilingual search by cascading a machine translation step before searching the index in its primary language.



In practice, search query translation usually involves a **translation memory** matching step before machine translation.

What is translation memory (TM)

A **translation memory (TM)** is a database, stores the source text and its corresponding translation in language pairs that have been previously translated.

Example (German-English) :

rasierwasser→ aftershave

kinder schokolade → kinder chocolate

patek philippe → patek philippe

haus laboratories → haus laboratories

game of thrones staffel → game of thrones series

morgenmantel damen → dressing gown womens

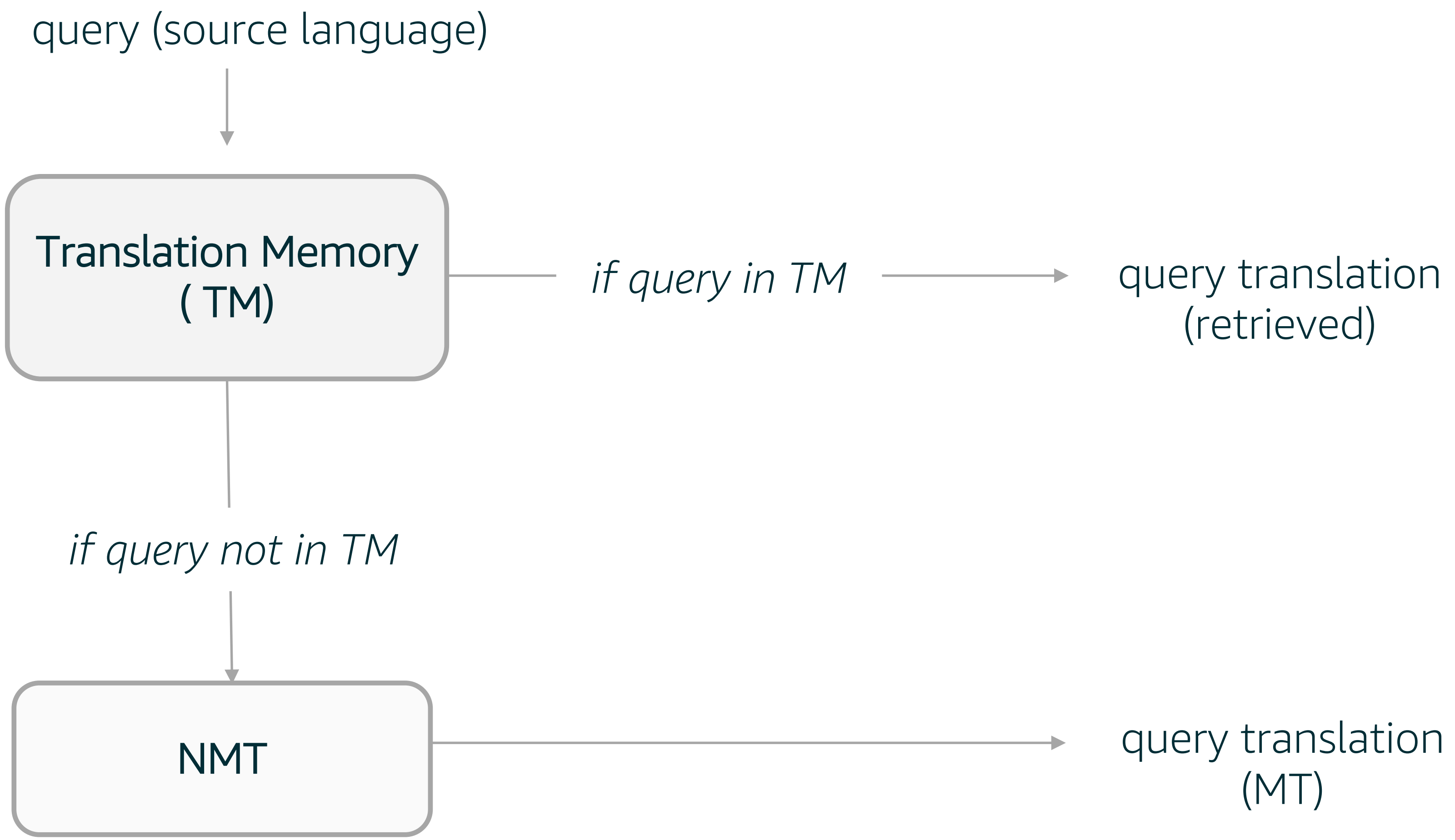
leinwände→ canvases

What can translation memory (TM) do?

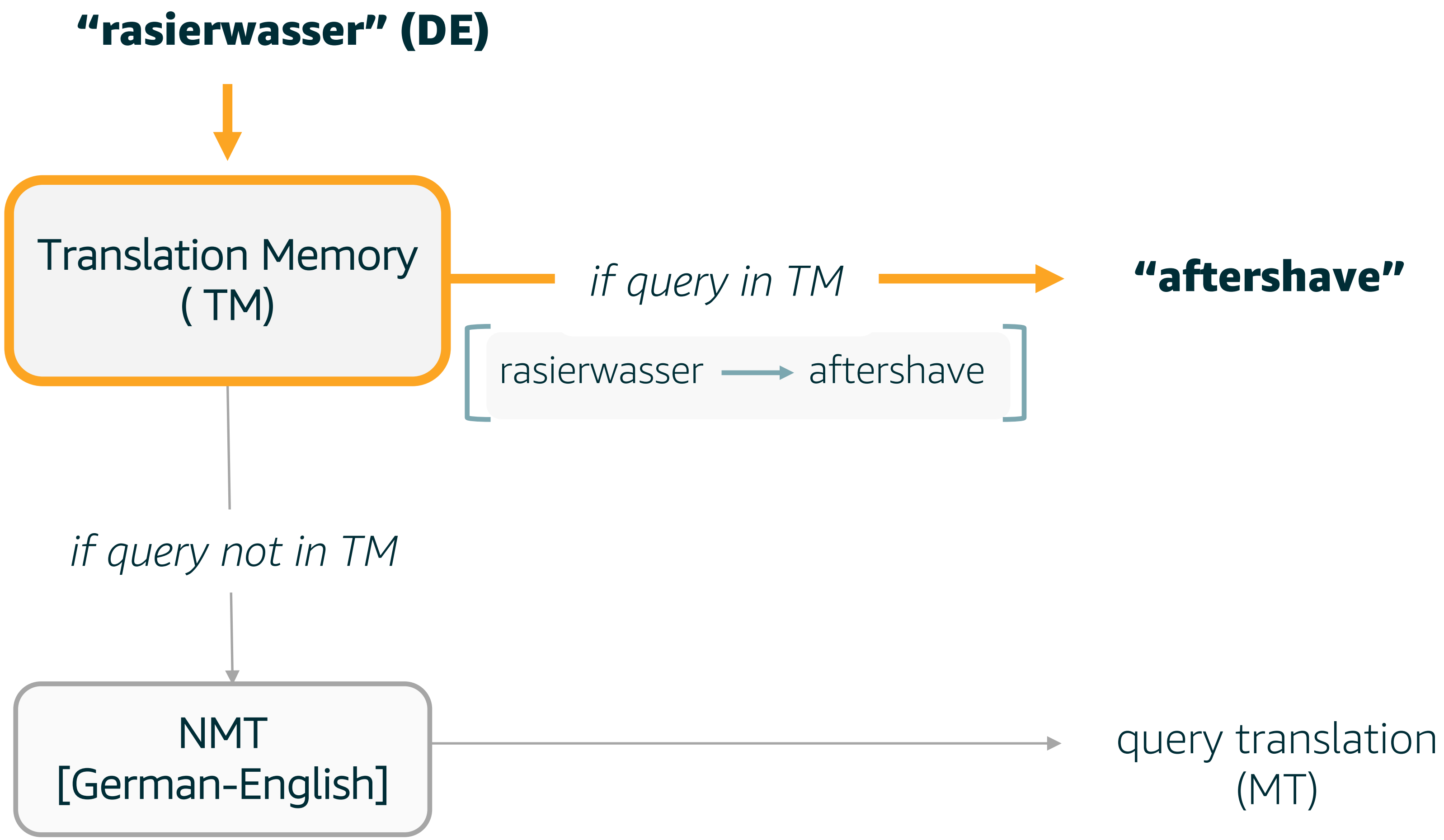
A translation memory (TM) can

- Effectively **enforce terminologies** for specific brands or products.
    - Although such issues can be mitigated through terminology constraint mechanism in the machine translation model, the turnover time to fix the translation would be unacceptable to the users and companies that expect an instant fix.
- **Reduce the computation footprint** and **latency** for synchronous translation.
- **Fix machine translation issues** that cannot be resolved easily or quickly without retraining/tuning the machine translation engine in production.

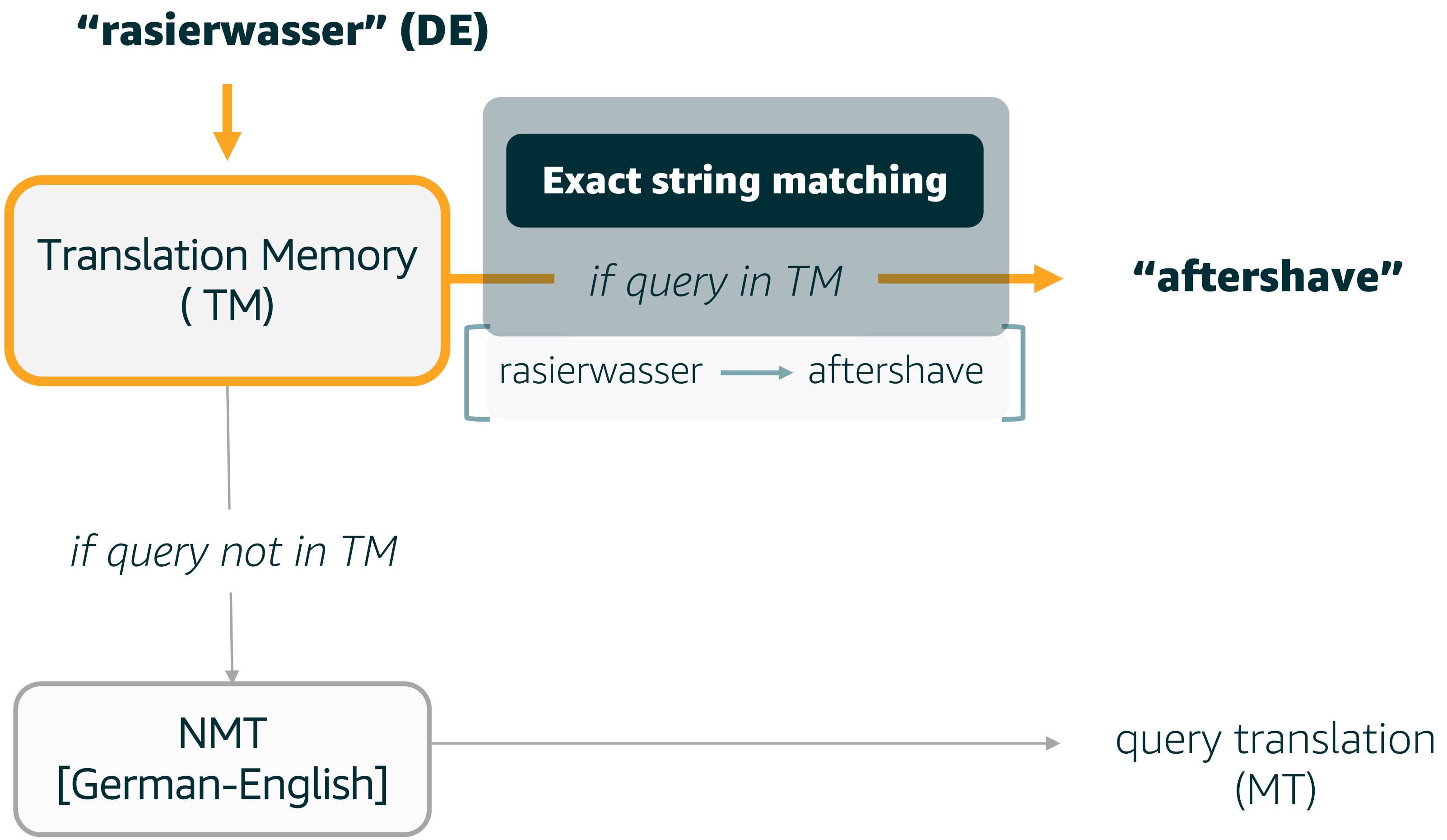
MT with translation memory (TM) - How do they interact in practice?



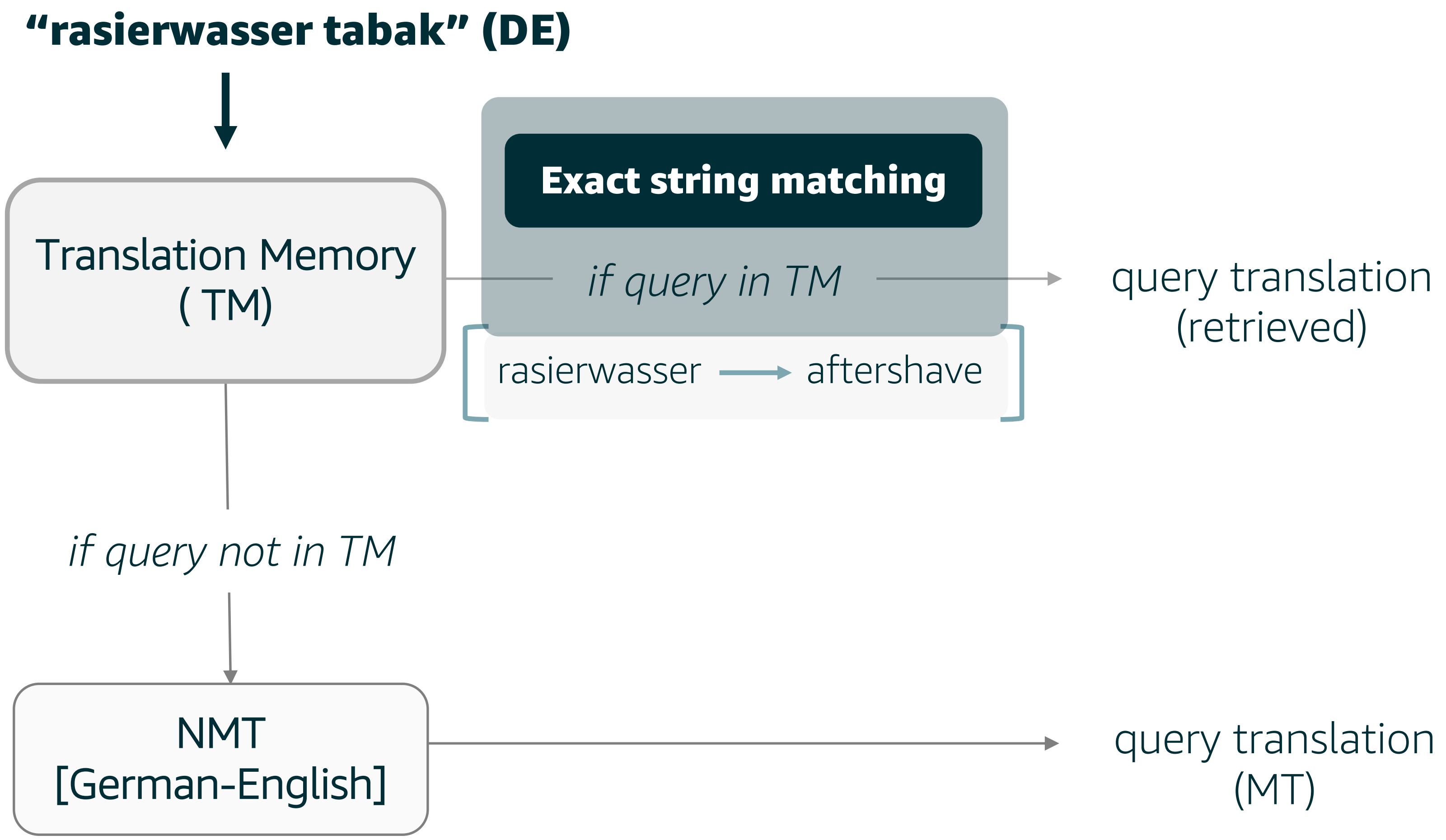
MT with translation memory (TM) - How do they interact in practice?



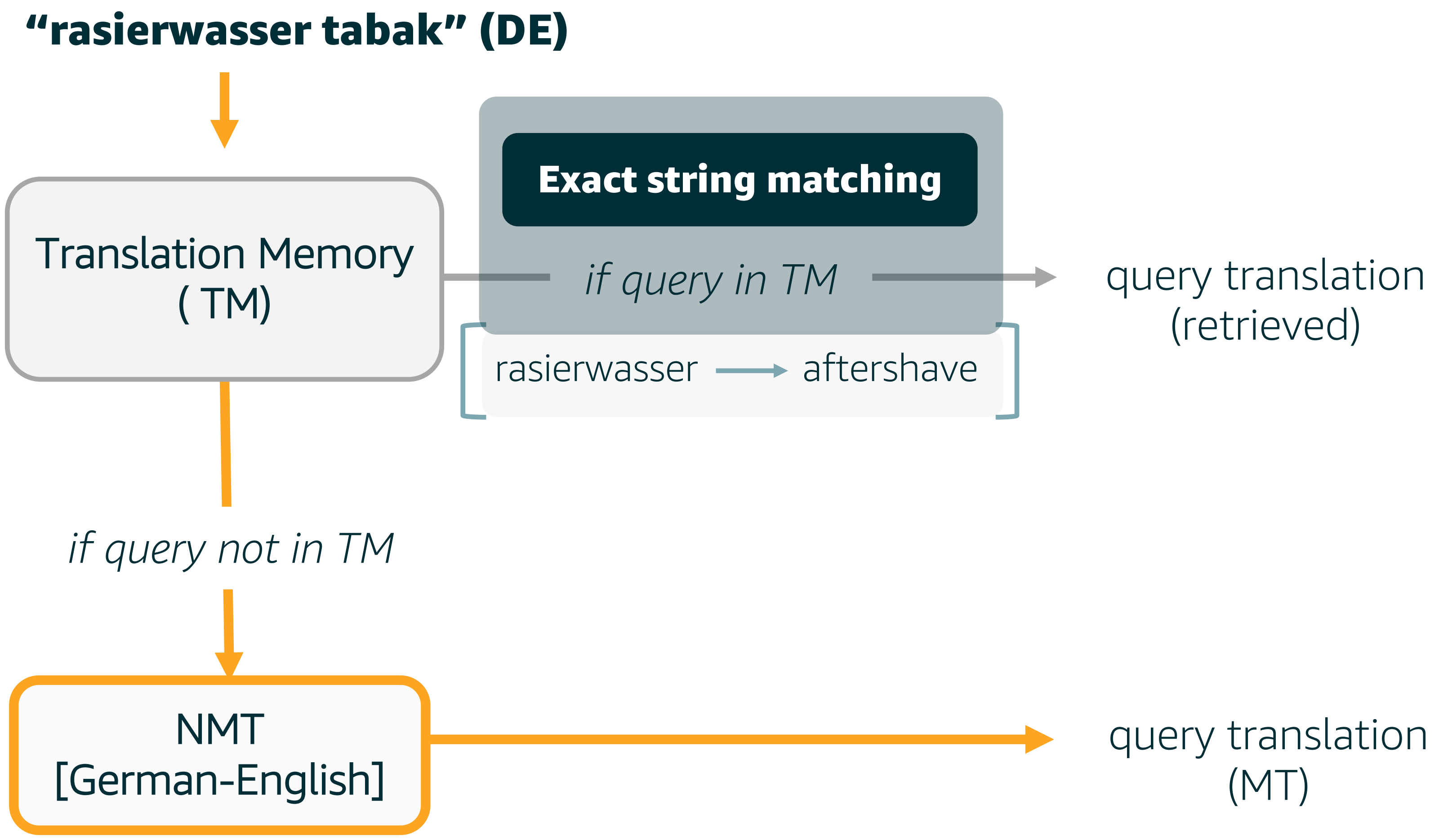
MT with translation memory (TM) - How do they interact in practice?



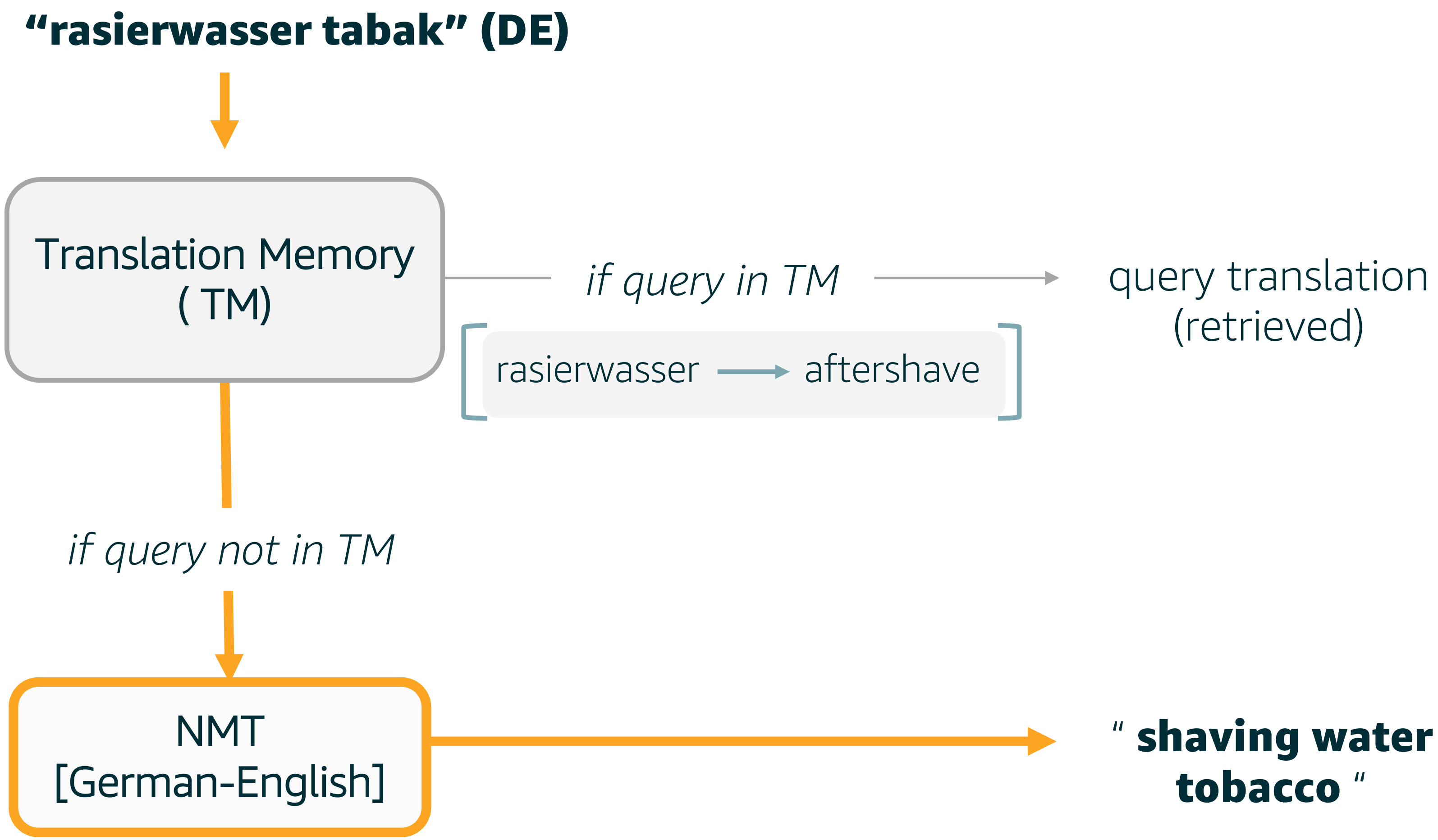
MT with translation memory (TM) - How do they interact in practice?



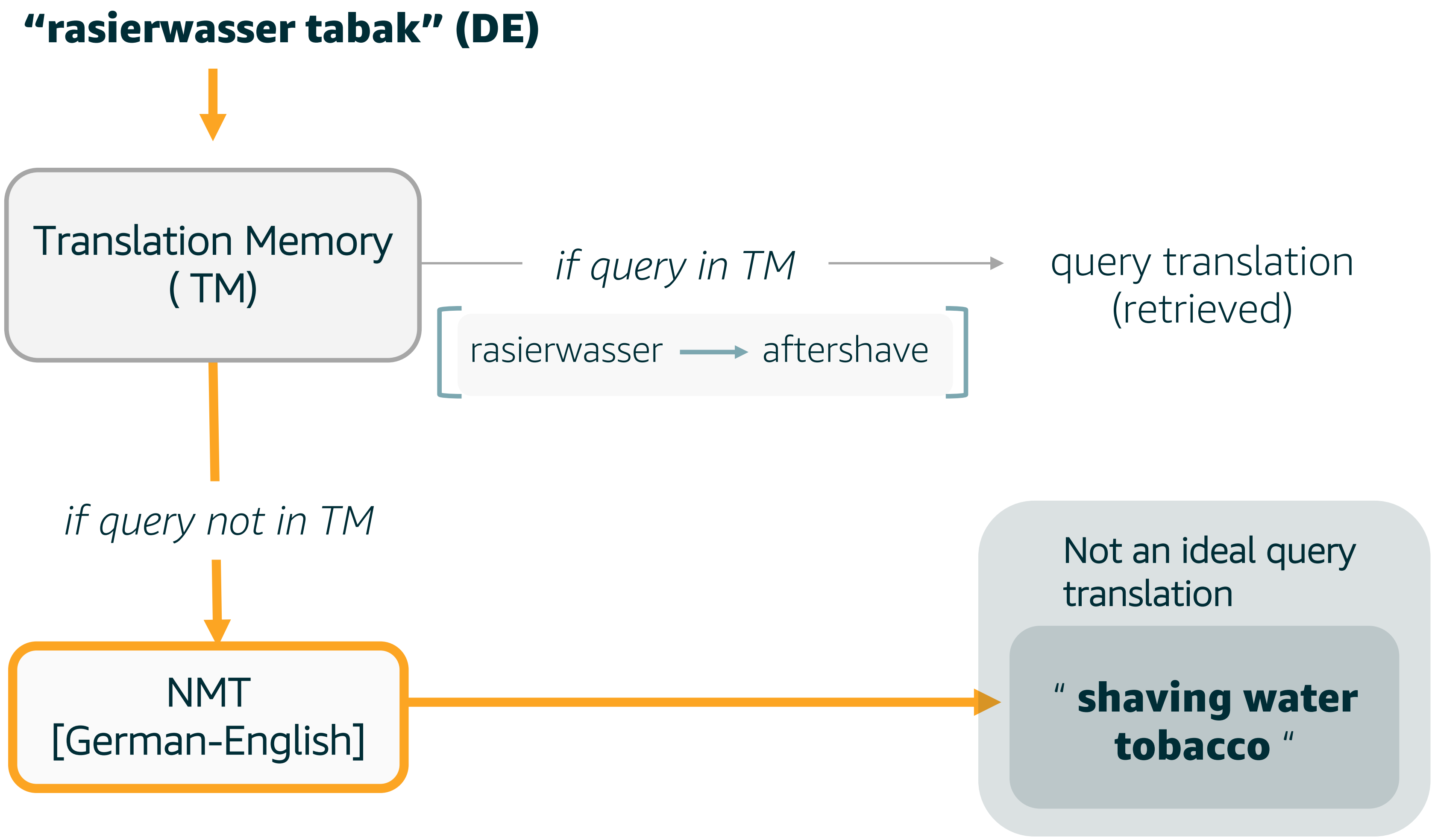
MT with translation memory (TM) - How do they interact in practice?



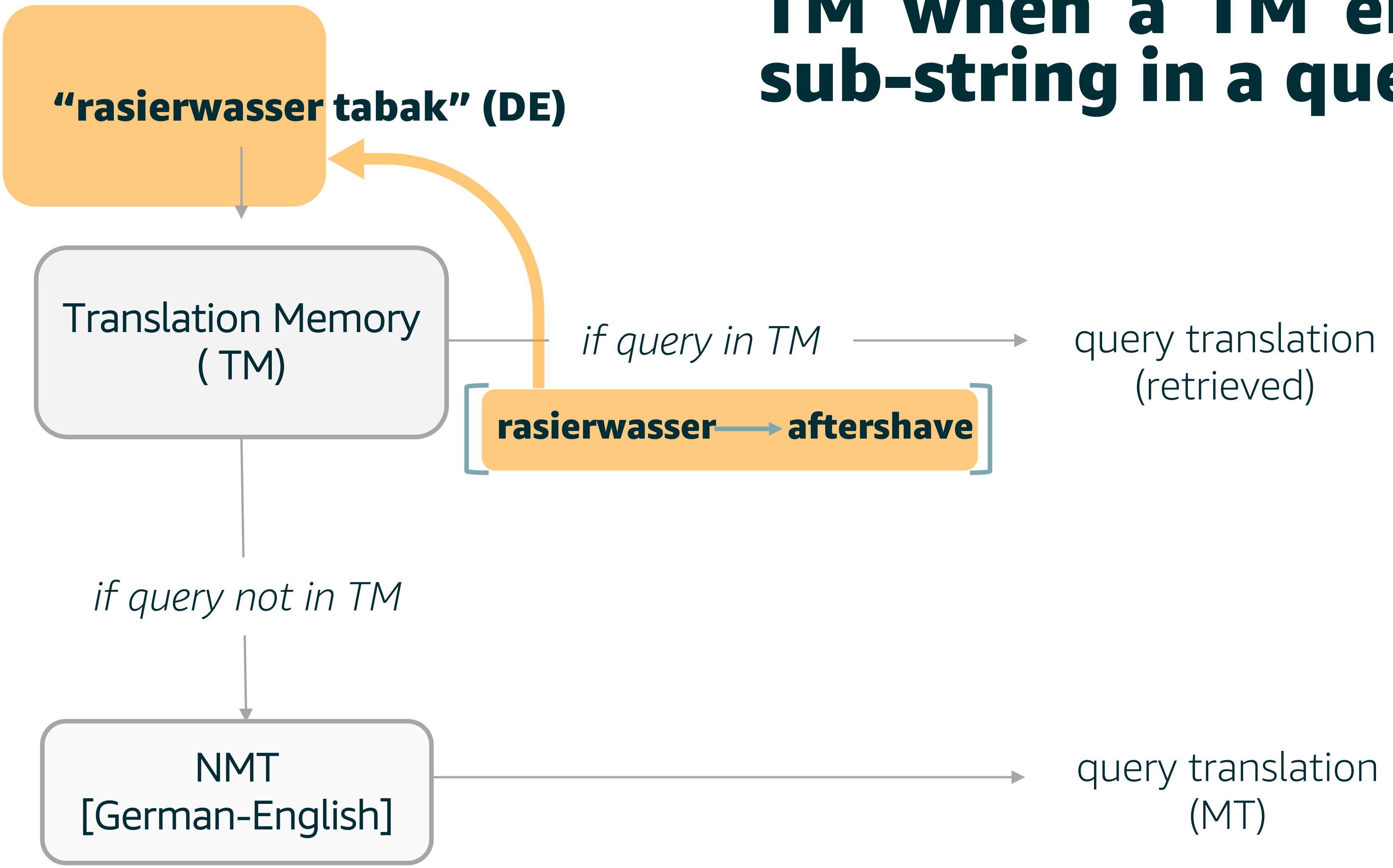
MT with translation memory (TM) - How do they interact in practice?



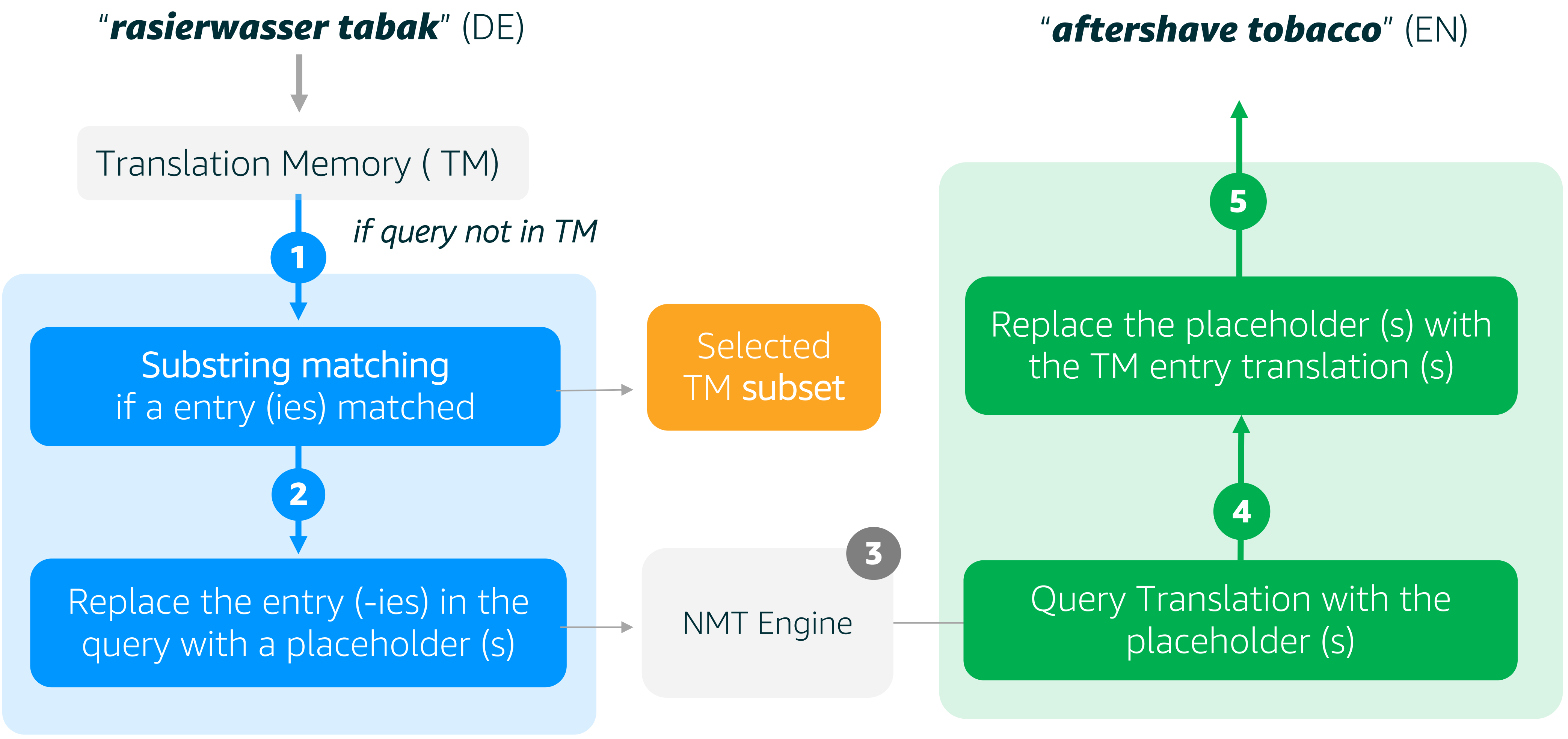
MT with translation memory (TM) - How do they interact in practice?



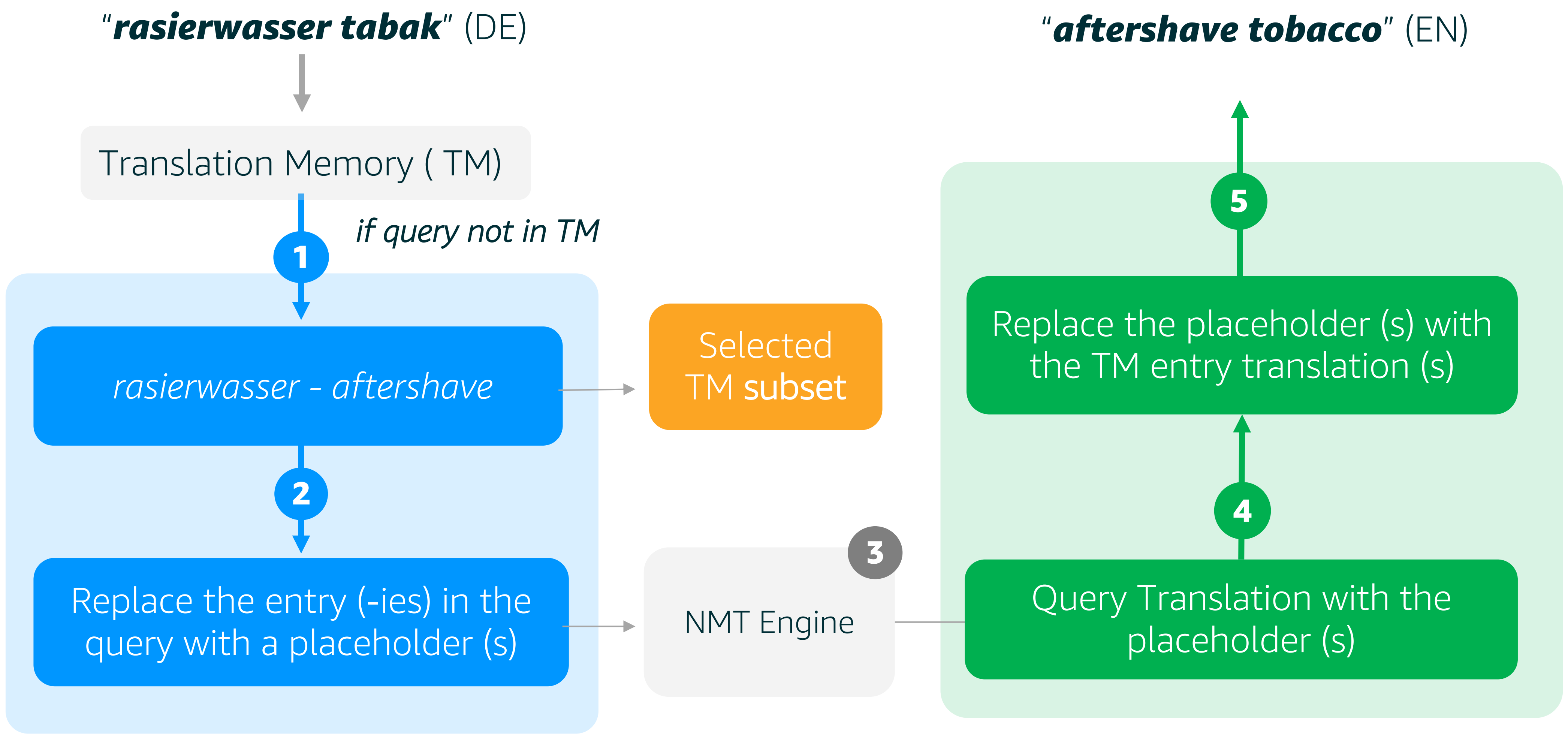
Can we improve MT using TM when a TM entry is a sub-string in a query?



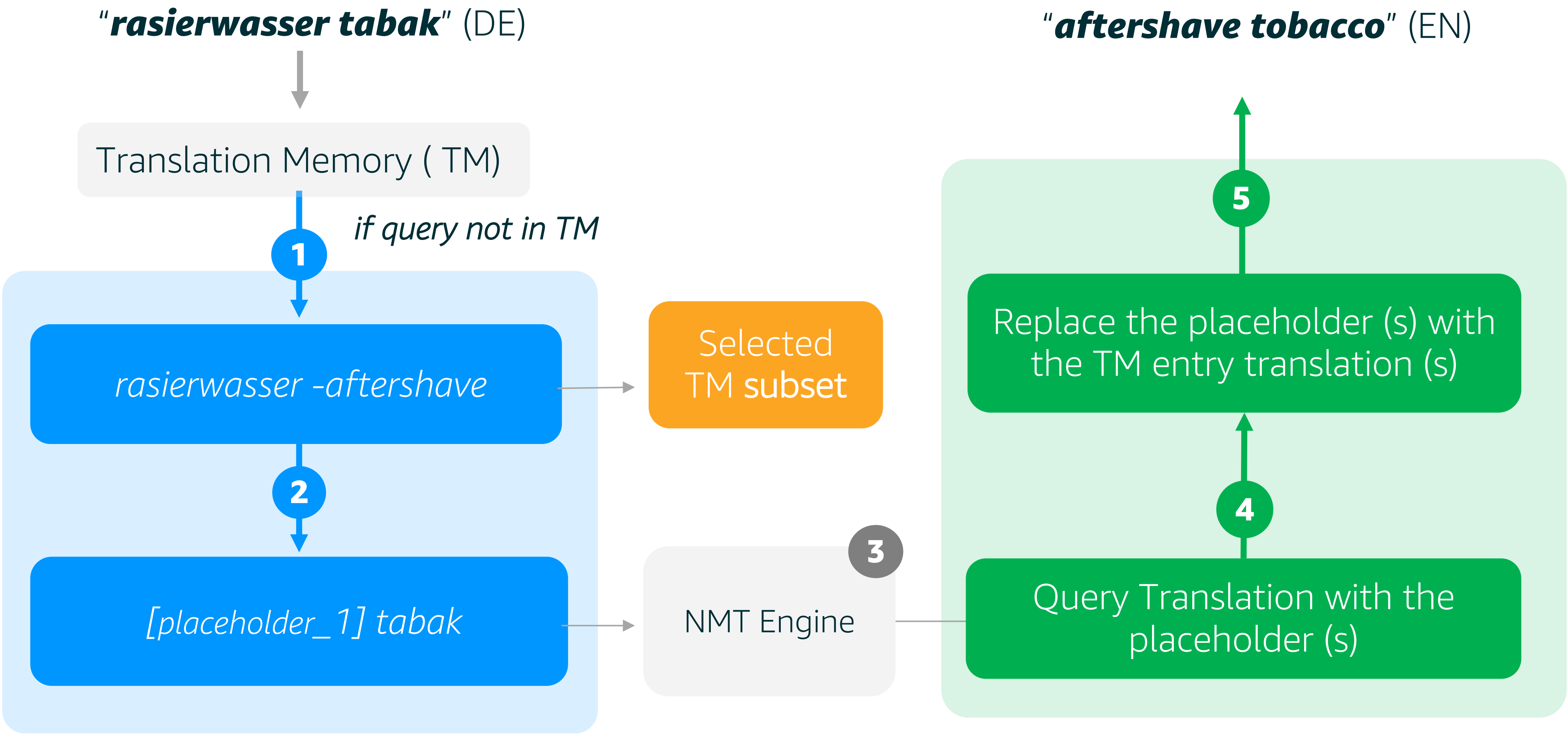
Improve neural machine translation (NMT) with selected translation memory



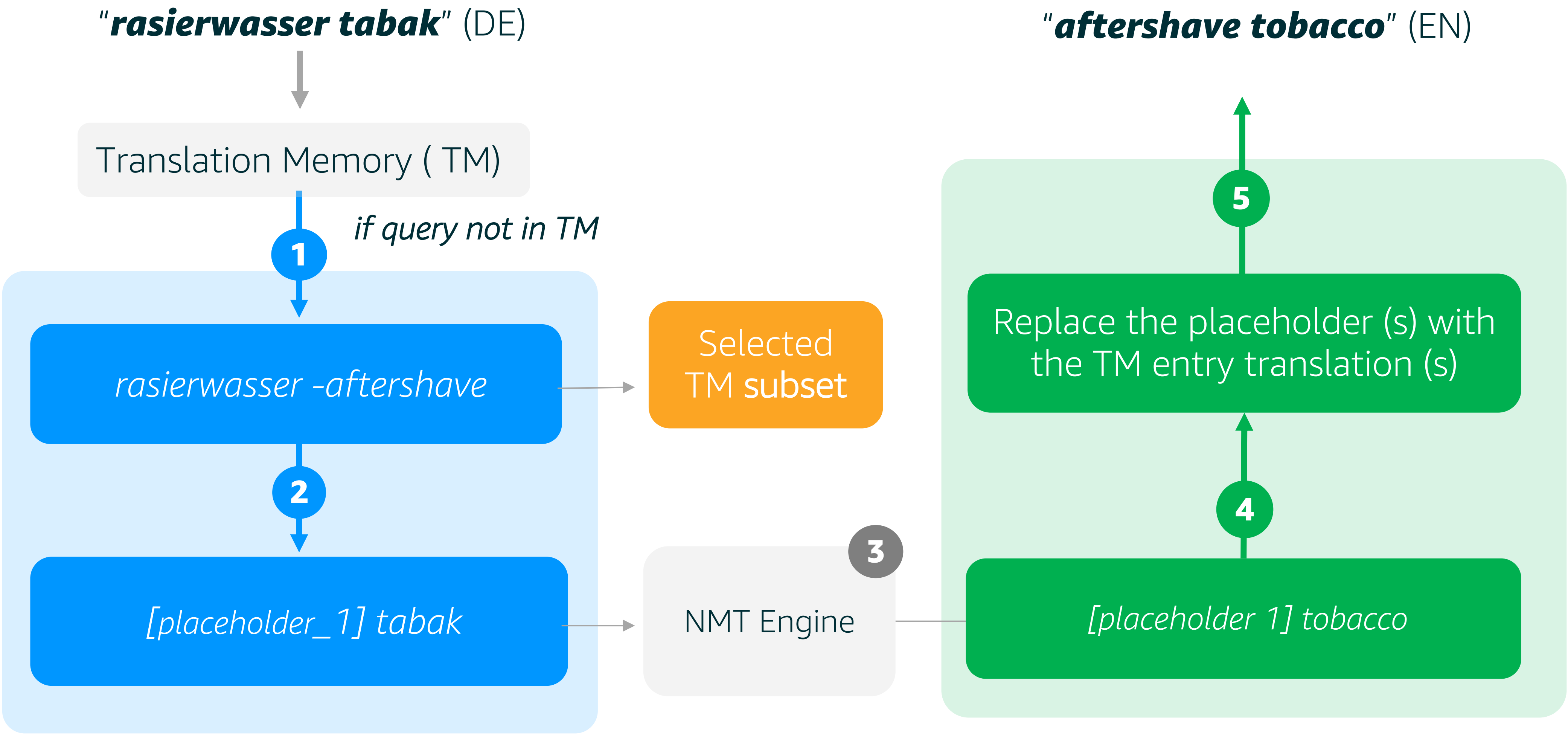
Improve neural machine translation (NMT) with selected translation memory



Improve neural machine translation (NMT) with selected translation memory



Improve neural machine translation (NMT) with selected translation memory



N-Gram Substring matching

- We implement a **back-off n-gram matching algorithm** that will match the translation memory entries in the source language to queries as sub-strings.

- Given a query, the query is first converted into **n-grams**, then we try to match the n-grams to the entries in the translation memory.

- We start the value of n as the number of the tokens in the query and then decrease the value of n for the n-grams until n = 1 or until we find a match in the memory. This way, we can aim at finding the **longest match**.

Query: "***rasierwasser tabak***" (DE)

N=2

2-gram: ["rasierwasser tabak"]

1-gram: ["**rasierwasser**", "tabak"]

[**rasierwasser** → aftershave]

Exploting Placeholder feature in modern Neural Machine Translation

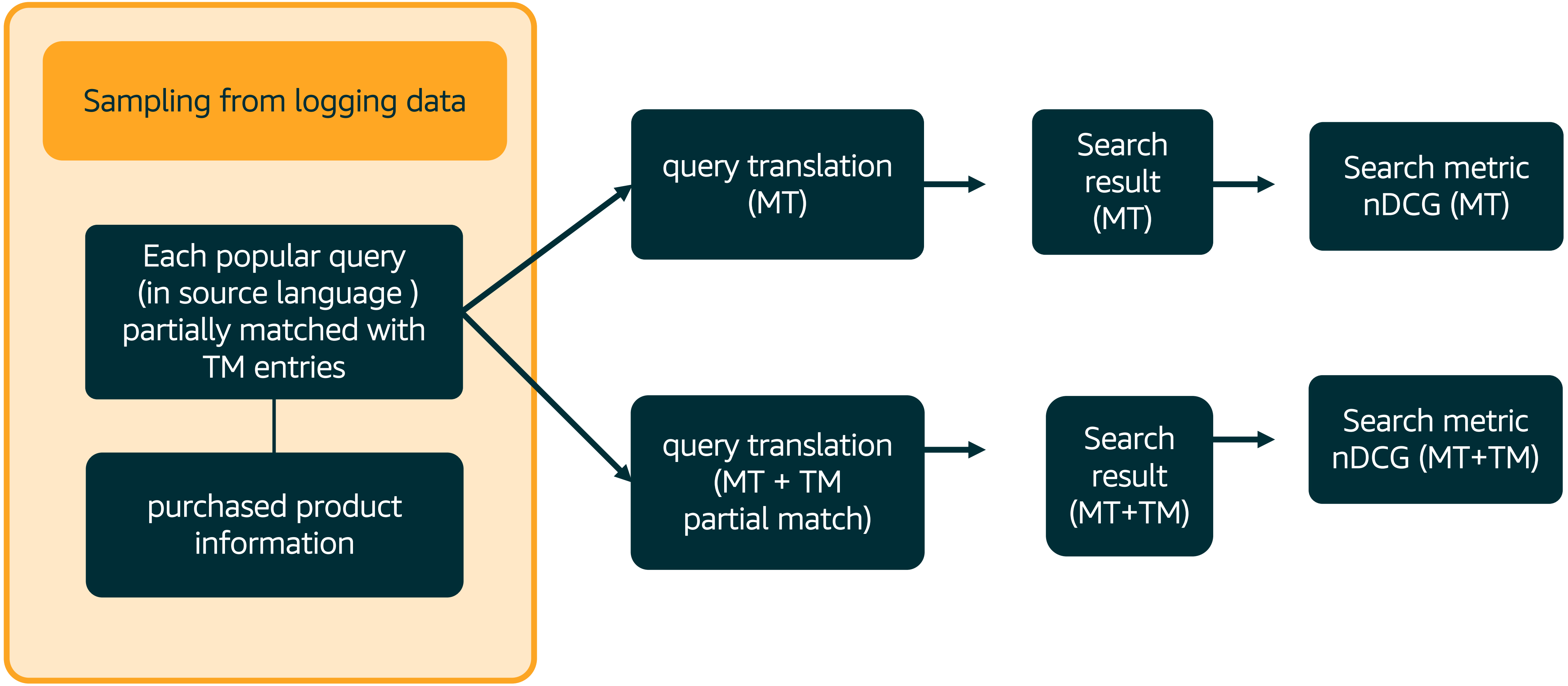
- We **augment** the neural machine translation (NMT) models with placeholder data during the training, so the NMT models can translate queries with placeholders and **keep those placeholders intact during the translation**.

- Those **placeholders are serialized** tokens e.g. placeholder 1, placeholder 2, which are part of the vocabulary used at inference time.

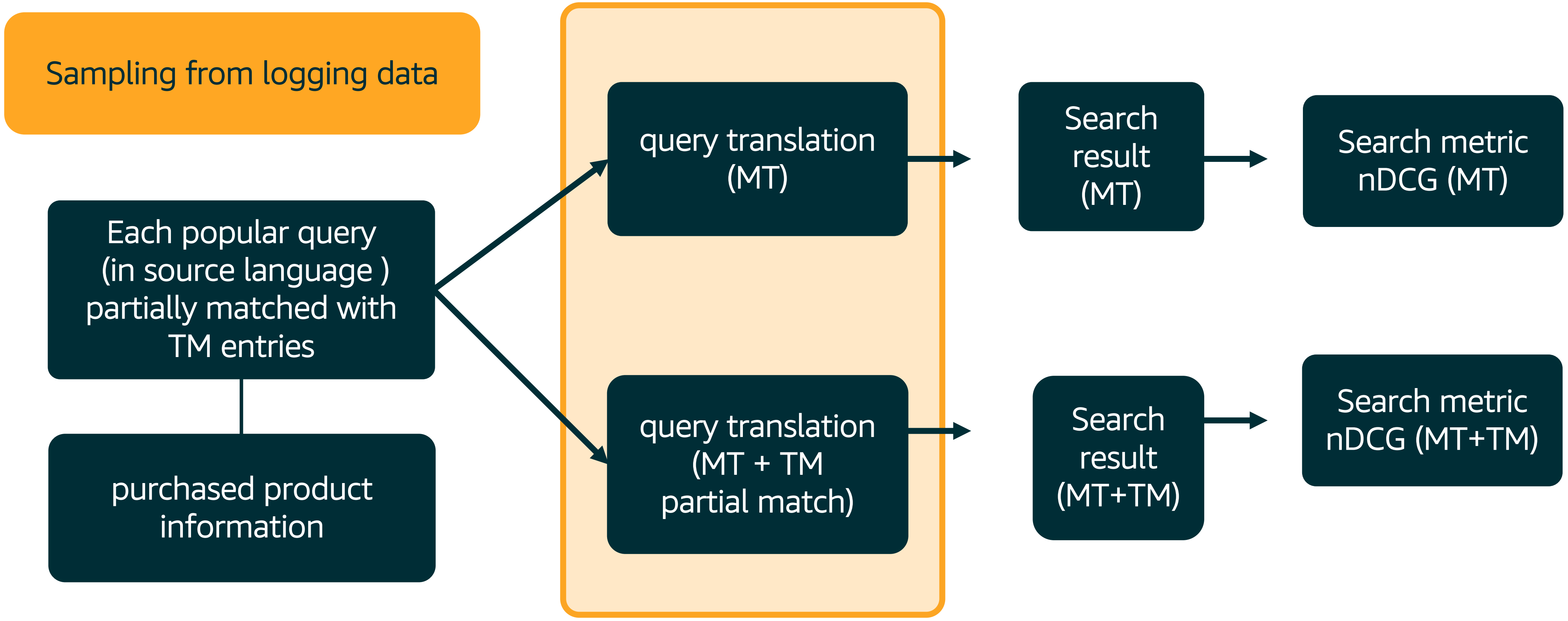
Why do we need to select a TM subset for substring matching?

- With substring matching , a TM entry can **impact a larger number of queries**. Therefore, a selection workflow is needed for validation.
- TM may have incorrect translation as well. The selection process also serves as sanity check.
- Need to select TM entries that will make bigger positive impact on query translation
  - TM entries translation that fit down stream search tasks better.
  - TM entries that **MT cannot translate well**.
  - **Terminology** translation

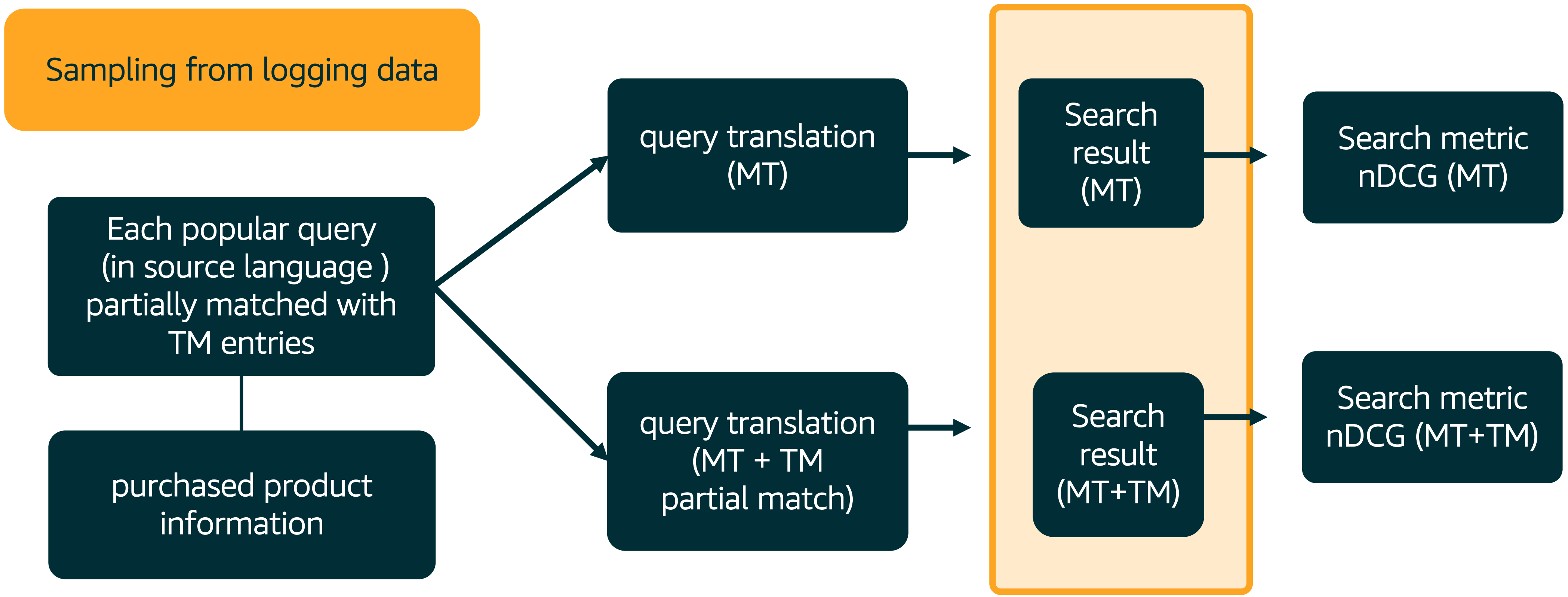
How to select an optimal translation memory subset ?



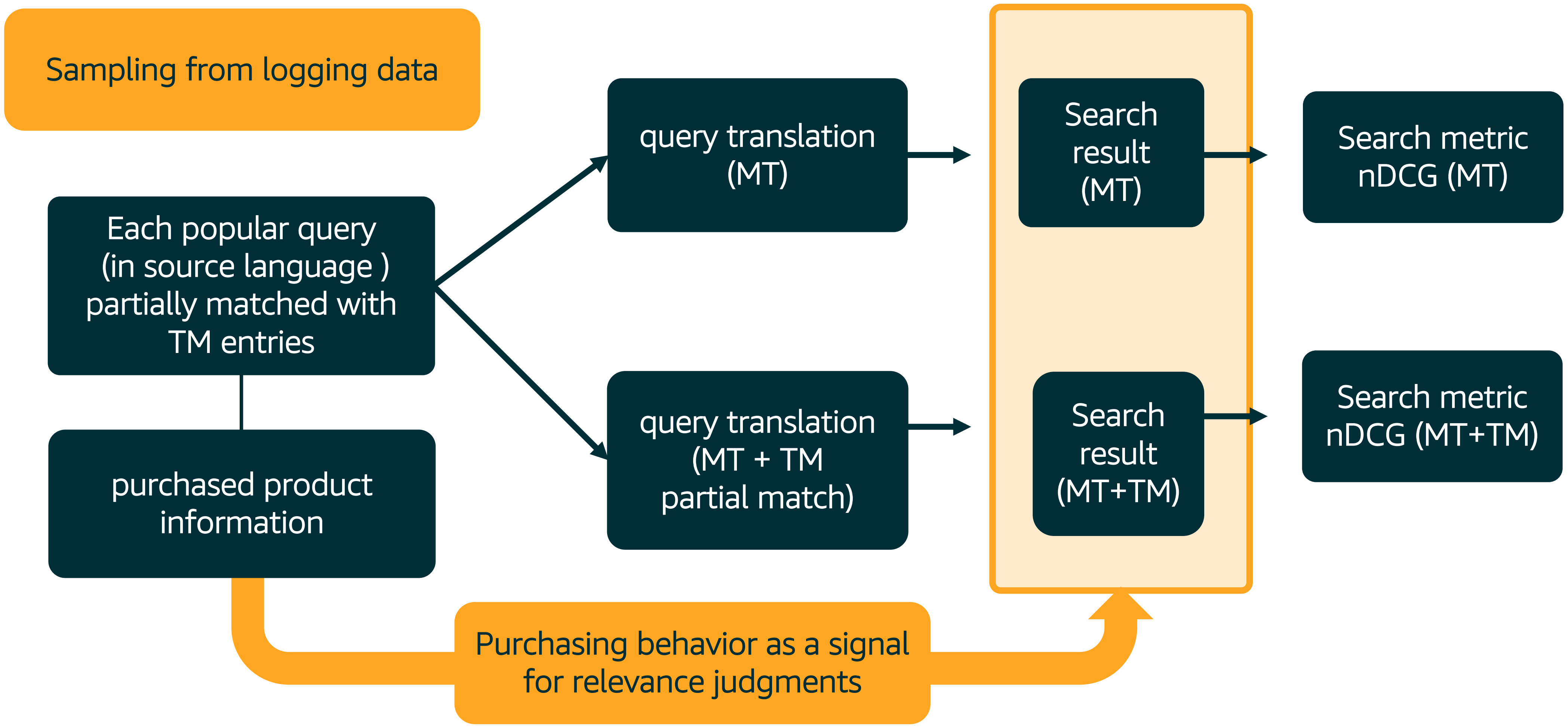
How to select an optimal translation memory subset ?



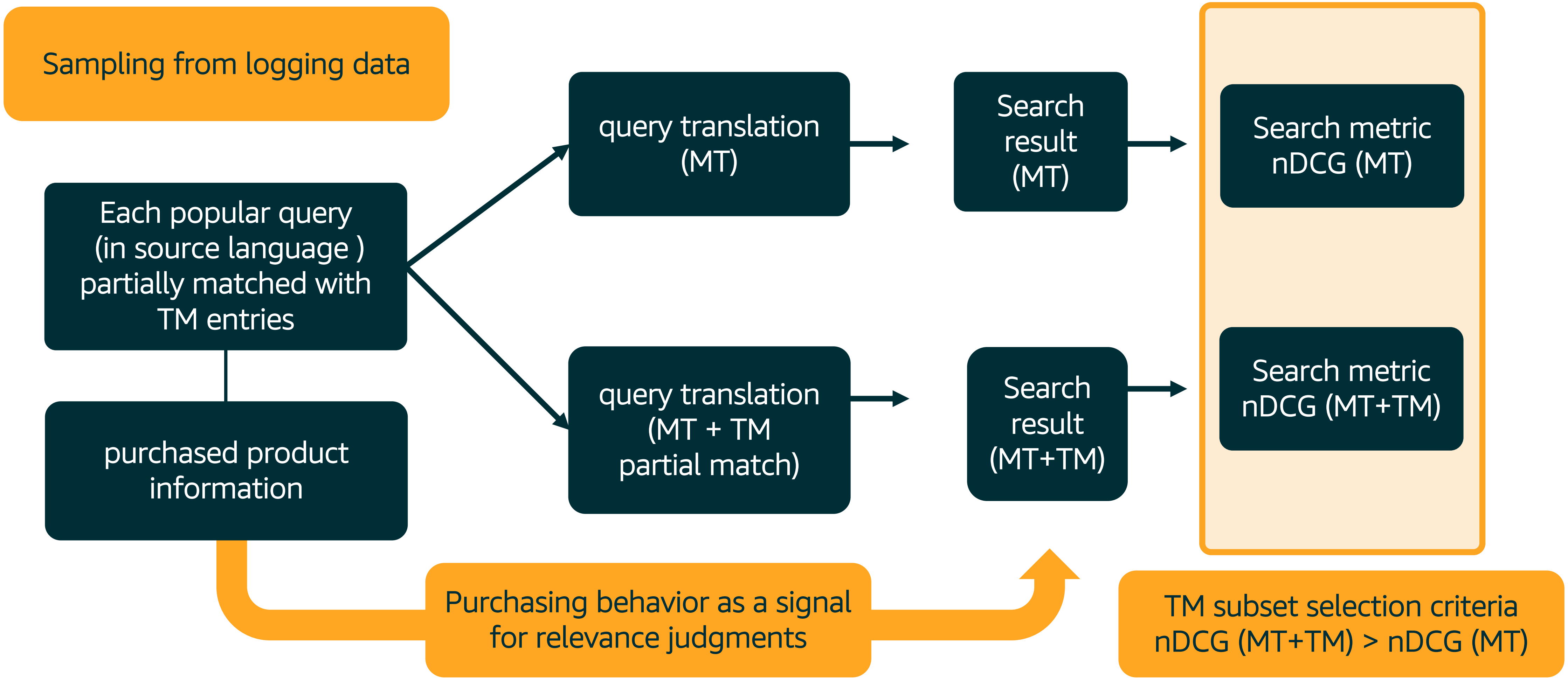
How to select an optimal translation memory subset ?



How to select an optimal translation memory subset ?



How to select an optimal translation memory subset ?



TM entry modification

- We also observe many **brand entries** overlap with **the common vocabulary of the source language** that usually **needs to be translated**:
  - For example, "*kinder*".
- This word is both a brand and a common word in German meaning *children*; when it refers to the brand it is expected to be preserved in the German-to-English translation.
- Therefore, if such entries exist in the TM, we suggest creating a frequent collocation *kinder schokolade* alone based on the query log and adding the new entry pair *kinder schokolade - kinder chocolate* to the translation memory before the subset selection.

Experiment - Setup

- We experiment with three language pairs:

  - German - English
  - Dutch – German
  - Portuguese – Spanish

- The subset of TM is selected using our proposed selection approach with nDCG@16 (normalized Discounted Cumulative Gain) as the search signal

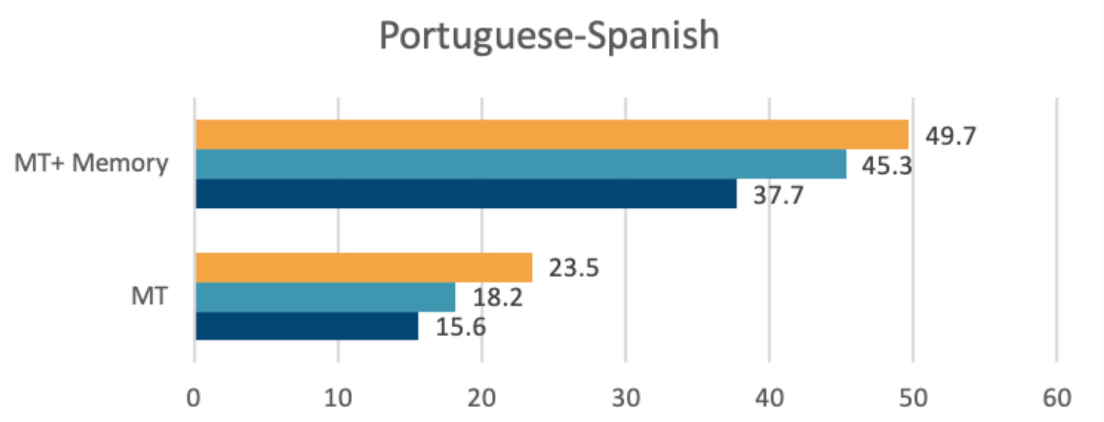
- Each language pair has 2500 test cases that are not used in the TM subset selection. Each query in the test cases can partially match a unique entry from the selected TM.

Experiment-Result

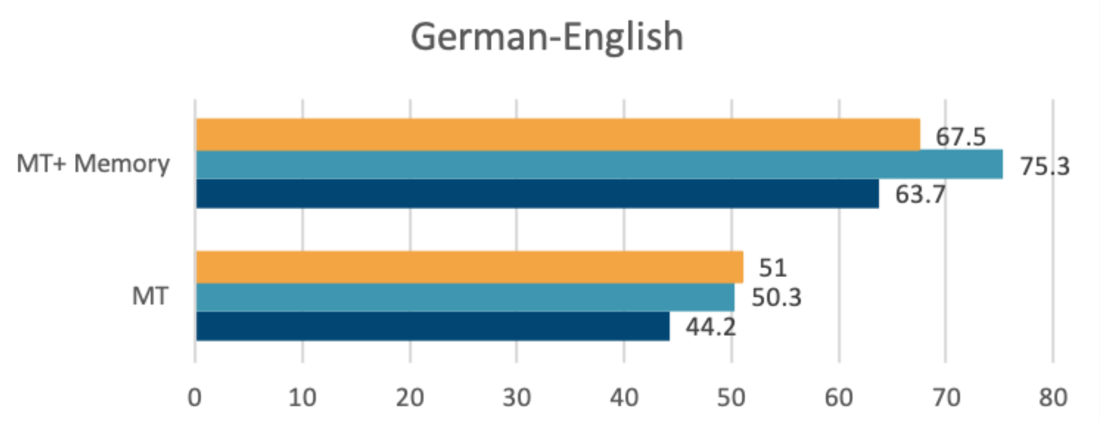
- We use the following search metrics to evaluate our solution.
  - **MAP** (Mean Average Precision)
  - **MRR** (Mean Reciprocal Rank)
  - **nDCG** (normalized Discounted Cumulative Gain)

- All the search metrics have been scaled from 0-1 to 0-100.

Experiment-Result

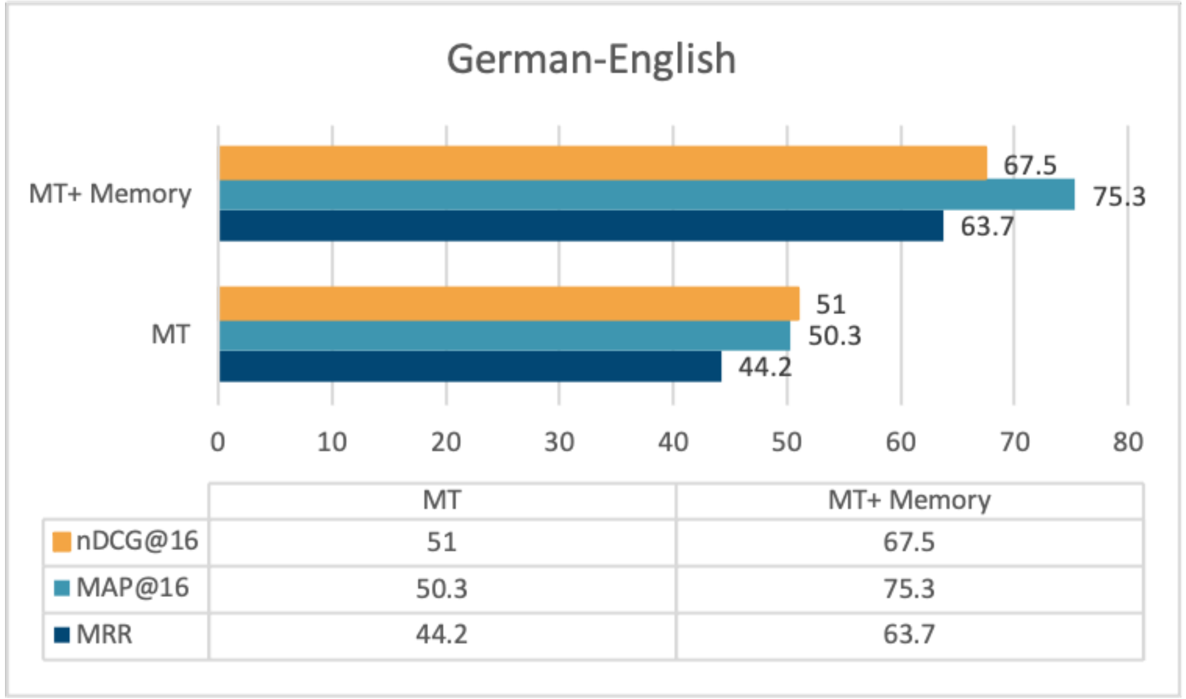


	MT	MT+ Memory
nDCG@16	23.5	49.7
MAP@16	18.2	45.3
MRR	15.6	37.7



	MT	MT+ Memory
nDCG@16	51	67.5
MAP@16	50.3	75.3
MRR	44.2	63.7

Experiment-Result



We have also conducted online experiments for the proposed approach with selected TM subset using search signal for:

- Portuguese queries on Amazon.es
- German queries on Amazon.com (US)
- Dutch queries on Amazon.de.

All three stores have seen increased order product sales (OPS) and improved user experience.

The following tables are showing the examples of the default MT query translation and improved MT query translation with the selected TM entry substitution.

Query in source language	Translation (MT)	Translation (MT + TM)	Selected TM entries (source-target)	
jurassic world lego sets günstig	jurassic world lego sets cheap	jurassic world legacy sets cheap	jurassic world lego	jurassic world legacy
happy hippos kinder chocolate	happy hippos kids chocolate	happy hippos kinder chocolate	kinder chocolate	kinder chocolate
freizeitkleider für damen weiß	leisure dresses for women white	casual dresses for women white	freizeitkleider für damen	casual dresses for women
game of thrones staffel 8	game of thrones relay 8	game of thrones series 8	game of thrones staffel	game of thrones series
inliner herren grösse 43	inliner mens size 43	roller blades mens size 43	inliner herren	roller blades mens
mitesserentferner set	mitesserremover set	blackhead remover set	mitesserentferner	blackhead remover
büromaterial mappe 1-12	office material folder 1-12	office supplies folder 1-12	büromaterial	office supplies
rasierwasser tabak	shaving water tobacco	aftershave tobacco	rasierwasser	aftershave
ordnungsbox gold	ordering box gold	storage box gold	ordnungsbox	storage box

Reference

[1] Lowndes, M. and Vasudevan, A. (2021). Market guide for digital commerce search.

[2] Caskey, S. P. and Maskey, S. (2013). Translation cache prediction.

[3] Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

[4]Kanavos, P. and Kartsaklis, D. (2010). Integrating machine translation with translation memory: A practical approach. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 11–20, Den- ver, Colorado, USA. Association for Machine Translation in the Americas.

[5] Luo, C., Lakshman, V., Shrivastava, A., Cao, T., Nag, S., Goutam, R., Lu, H., Song, Y., and Yin, B. (2022). Rose: Robust caches for amazon product search. In *The Web Conference 2022*.

Reference

[6] Post, M. and Vilar, D. (2018). Fast lexically constrained decoding with dynamic beam allo- cation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

[7] Susanto, R. H., Chollampatt, S., and Tan, L. (2020). Lexically constrained neural machine translation with Levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

[8] Wang, K., Gu, S., Chen, B., Zhao, Y., Luo, W., and Zhang, Y. (2021). TermMind: Alibaba's WMT21 machine translation using terminologies task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 851–856, Online. Association for Compu- tational Linguistics.

[9] Ailem, M., Liu, J., and Qader, R. (2021). Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL- IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Thank you for your attention!