Improve Machine Translation in E-commerce Multilingual Search with Contextual Signal from Search Sessions

Bryan Hang Zhang[1,†], Taichi Nakatani[1,†], Stephan Walter[1], Amita Misra[1] and Elizabeth Milkovits[1]

[1] *Amazon*

Abstract

Over a period of years, search engines have become adept at understanding and providing relevant results for short user generated queries for monolingual search. However, the brevity of search queries can be a limitation for cross-lingual e-commerce search. Previous studies have demonstrated that discourse-level context information can improve machine translation (MT) for document translation but there is no well-defined context regarding MT for query translation. Therefore, in this study, we aim to improve MT for search by incorporating contextual signals from search sessions. Our first step is to explore and categorize two types of contextual queries from search sessions: those with content variations and those with spelling variations. We then propose an innovative approach to derive bilingual training data from search sessions and incorporate the session queries as contextual signals. Using this data, we augment the training data to improve MT. Our initial experimental results demonstrate that augmenting the training data with content variant session queries as context can enhance MT for query translation. Overall, our study provides insights into how contextual information from search sessions can be leveraged to improve machine translation in multilingual e-commerce search.

Keywords

machine translation, cross-lingual retrieval, e-commerce multilingual search

CEUR-WS.org/Vol-3589/paper_6.pdf

1. Introduction

Multilingual search capability is essential for modern e-commerce product discovery [1, 2]. Localization of e-commerce sites have led users to expect search engines to handle multilingual queries. Recent proposals such as multilingual information retrieval and product indexing has gained traction with neural search engines [3, 4, 5, 6, 7]. However, many e-commerce search indices are still built on monolingual product information and multilingual search is supported though query translation [8, 9, 10, 11, 12, 13].

Query translations are a key component in large multilingual e-commerce stores because it allows users to find product information written in languages different from the language of the query. Given a query in the source language, it uses its translated form as the input for the search engine to retrieve documents in the target language. Search engines typically have

eCom'23: ACM SIGIR Workshop on eCommerce, July 27, 2023, Taipei, Taiwan

[†] These authors contributed equally.

bryzhang@amazon.com (B. H. Zhang); taichina@amazon.com (T. Nakatani); sstwa@amazon.de (S. Walter); misrami@amazon.com (A. Misra); emilko@amazon.com (E. Milkovits)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

preferred word choices and collocations based on users' query patterns [14, 15], and previous studies have demonstrated that better translation quality improves retrieval accuracy [16, 17, 2].

Typical modern neural machine translation models for query translation (Search MT) are trained on bilingual query data for domain adaptation. User-generated queries are short and have limited textual context in query texts, which can pose difficulties for search MT to learn word senses and choices as well as other linguistic aspects sufficiently during training. Meanwhile, incorporating context in the neural machine translation is studied extensively for document-level machine translation (Document MT). Previous studies show that neural machine translation can distinguish and learn from the discourse history when the source texts of the training data is extended with document-level context [18]. However, little attention has been received on exploring and incorporating contextual signals for Search MT. Unlike Document MT where neighboring sentences in the document of an input sentence can serve as a natural source of context, it is not as straightforward to define the context for an input query of Search MT. Furthermore, context-aware MT systems usually require the context as part of the input at inference time; In an industry setting, it is preferred not to modify the run-time input query of the search MT as it adds further complexity such as latency to the large search ecosystem. Therefore, in this paper, we present a pilot study that uses queries from search sessions as contextual signals to improve search MT which does not require changes in run-time input and output setups. We first categorize two types of contextual queries from search sessions: content variant queries and spelling variant queries, and we propose using Levenshtein Edit Distance [19] as a soft approximation to differentiate the two types. We then propose an innovative approach to derive bilingual training data from search sessions and incorporate the session queries as contextual signals. Using this data, we augment the training data to improve Search MT. Our initial results show augmenting the training data with context session queries that are mainly content variations can improve the search MT for English-German by +0.7 BLEU.

The contributions of this paper are: (1) Explore and analyze the user-generated query data from search sessions in the e-commerce multilingual search; (2) Propose an approach to incorporate queries from search sessions as contextual signals in the bilingual query data; (3) Propose a method to augment search MT training using the bilingual query data extended with contextual signals.

2. Session-based queries in e-commerce multilingual search

Users are provided with the option to search and browse products in their preferred language which differs from the primary language of the store. For example, users can query in English in the German store where the primary language is German. Given a query from a search session of a user, the previous and the next queries of the current query can serve as **contextual queries** and provide more information to the **current query**. Based on our analysis and observation [1], we propose two main categories of contextual query(ies) for a given current query, namely content variant queries and spelling variant queries.

[1] Refer to section 4.1 for more details on stats of the Search Session Data for the analysis

2.1. Content variant queries

Semantically-related queries are contextual queries that are semantically related to the current query. For example, *vanilla extract for baking* -> *vanilla* or *guitarras elétricas yamaha* -> *yamaha pacifica* (Portuguese); *accessoire ordinateur* -> *tapis de souris, bmw 328i 2011 grille calandre* -> *grille calandre* (French). These queries are all topically related, likely stemming from a similar shopping intent.

Multilingual queries are semantically identical or similar to the current query but partially or entirely in the primary language, which are particularly common and unique in the e-commerce multilingual search. For example, *laserdrucker multifunktionsgerät* (German) -> *laser printers*, *tiroir plastique organisateur* (French) -> *plastic drawer*; *abajur star wars* (Portuguese) -> *star wars lamps for adults.* This phenomenon can be from users' curiosity for product discovery with queries in both the preferred language and primary language of the region, or search results may be unsatisfactory in one language so they may try another language, but the reason for this behavior is not in the scope of this study.

2.2. Spelling variant queries

While topically-related and multilingual contextual queries are mostly content variants of the current query, there are also a number of contextual queries that are spelling variants of the current query. For example, *akubormaschine->akubohrmaschine* (German), *adidad tshirt* -> *adidas tshirt.* The majority of such cases come from users' spelling correction behavior.

3. Incorporating Queries from Search Session as Context for Search MT Training

User-generated queries from search sessions are related to each other as discussed in Section 2, the previous and the next queries of the current query can serve as context to provide more information for the current query. Therefore, we propose to restructure such query data and use it to augment the training data for MT training.

Given a bilingual query set Q collected from query log, $Q = (p_1, p_2 ... p_n)$ where p_1 is a bilingual query pair with contextual queries from the search session; $p = (q_{prev}, q_{src}, q_{next}, q_{tgt})$ where q_{src} is the current query in the source language, q_{prev} is the previous query in the source language from the same search session of a given user, q_{next} is the next query from the same search session of the given user, q_{tgt} is the current query in the target language generated by the Search MT in production.

We restructure each query data point p_i into two bilingual training data pairs p_i^1 and p_i^2 as Figure 1. $p_i^1 = (q_{prev} + q_{src}, q_{tgt})$ concatenates the current query and previous query as one query in the source language with the query in the target language, while the other pair $p_i^2 = (q_{src} + q_{next}, q_{tgt})$ concatenates the current query and the next query as one query in the source language with the query in the target language.

We propose to use this restructured data $Q' = \{p_1^1, p_1^2, ... p_n^1, p_n^2\}$ from search sessions to augment the training data for training Search MT. Intuitively only content variant queries from

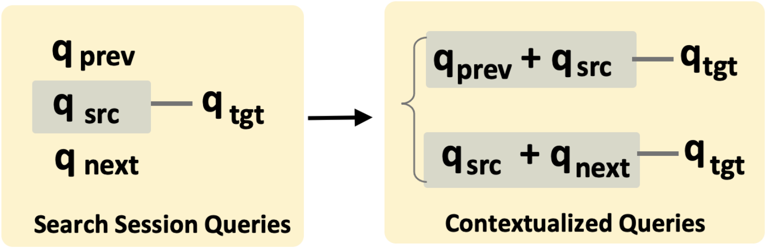

Figure 1: Restructure search session queries as contextual signals in the bilingual query data

sessions (in the source language) can provide more contextual information that is beneficial for MT training. Therefore, in cases where there are more spelling variant queries, we propose to use edit distance metric such as **Levenshtein Edit Distance** [19] between the previous (next) query and current query to separate spelling variant and content variant contextual queries.

4. Experiment

4.1. Experiment Setup and Evaluation

Language Pairs and Stores: We select three language pairs from three stores for our experiment: engb-dede (English in German store), frca-enca (Canadian French in Canadian store) and ptbr-enus (Portuguese in US store)

Search Session Data Sampling: For each language pair, we collect the user-generated query in the source language which is the preferred language in a given store (e.g. queries in English from the German store) and its query translation returned from MT which have been used for the downstream search tasks (e.g query translation in German for the German store as the search index is in German) if the query translation has results in purchases or a click(s) to ensure the query translation quality. Additionally, we collect the previous and next queries of the source query (also referred as ”current query” in this paper) from the same search session of a given user. Previous and next queries are in source language.

Test Data: Query translations from the Search MT are used for the downstream search tasks, therefore, we create test sets using the workflow proposed by the study [20] to evaluate both the MT translation quality and the search performance. We first sample the query data from historical search traffic that are generated by customers in the target (primary) language of store [2]. Empirically, we sample queries from the top 30%, bottom 30% and the middle 40% in frequency bins to reflect the distribution of user traffic. To allow computation of traditional relevance metrics, we aggregate the purchase product IDs associated with the queries if they are available. Given queries collected in the target language, language experts translated these queries to the source language.

[2] The search index is built on the primary language of the store.

Lang Pair	Context Queries	Data Size	Mean (Lev)	Std (Lev)
engb-dede	prev	200K	4.90	12.01
	next	3 million	4.04	7.67
frca-enca	prev	2 million	1.47	1.65
	next	3 million	3.30	4.91
ptbr-enus	prev	2K	3.26	9.03
	next	30K	4.33	5.83

Table 1
Mean of Levenshtein distance between previous (prev)/next queries and current queries across language pairs

The test set is comprised of 4000 queries (as reference query translation) per store (e.g. German store), and each query is translated into their respective language pairs (e.g. German queries are translated to English for the engb-dede langauge pair). Purchased product IDs associated with these queries are additionally stored, and they are used as a proxy to search relevance labeled by human annotators. A previous study has shown that purchases are useful proxies to human relevance annotations[21]; We use the logarithm of the frequencies of purchased products as the relevance score.

Evaluation Metrics - MT and Search: We use MT quality metrics BLEU[3], COMET [23] and chrF [24] to evaluate the query translation quality. We use normalized Discounted Cumulative Gain (nDCG), Mean Average Precision (MAP), Precision and Recall [4] to evaluate the search performance of query translations. We set K to 16 for the top-K search results, using the top-16 products in the search results to compute nDCG@16, MAP@16, precision and recall@16.

4.2. Training data

Generic Data: We obtain a large quantity of general news and bilingual web data of over 200 million lines for frca-enca and engb-dede language pairs, and over 10 millions lines for the ptbr-enus language pair. This data is used to train a generic (out-of-domain) MT model.

Query Data with Prefix Context: Using the Search Session Data described in Section 4.1, we restructure query session data as described in section 3. If a previous or next query of the current query exists, we concatenate the previous query with the current query as the new source query, and the same procedure is also applied to the next query; we use the query translation returned from the Search MT in production as the query in the target language for training. We sample approximately 4 million lines for language pairs of engb-dede and frca-enca respectively. For the ptbr-enus language pair, the data set is much smaller at about 30

[3] SacreBLEU version 2.0.0 [22]
[4] Both the nDCG@16, MAP@16, precision, recall and chrF are scaled to 0-100 for computation convenience

thousand lines due to smaller volume of traffic.

Query Data with Selected Prefix Context: We further create another data set only keeping data whose contextual query (previous or next query) and current query has Levenshtein Edit Distance[5] more than the mean of the data sample as shown in Table 1 because these contextual queries are less likely to be the spelling variants of the current query, which can be beneficial for model training. This data is comprised of 3 million lines for engh-dede and frca-enca, and 15 thousand for ptbr-enus.

Query Data: (i) Human translated query data and (ii) synthesized query data generated by back-translation are used to fine-tune the generic MT model. (iii) Bilingual queries from the search session data (only source-target queries) are also used, with 200K lines for frca-enca and engb-dede and 1K for ptbr-enus. Human translated query data is comprised of queries translated from the source language to the target language. For back-translation, queries in the target language (e.g. German queries from the German store) are translated using a MT model trained on the reverse language pair (translating German queries to English). Data filtering is applied during data collection for back-translation; Only queries searched with a frequency greater than one were considered to reduce noise. A total of 120K human translations and 5 million back-translated queries were used for each language pair.

4.3. Machine Translation (MT) models

For model training, we use a transformer-based architecture [25] having 20 encoder and 2 decoder layers with the Sockeye MT toolkit [26]. In our experiment, for each language pair, we train three search machine translation systems namely M_0 (baseline), M_1 (full context), and M_2 (selected context). For each model, we first train a generic machine translation system then fine-tine this generic machine translation system using domain-specific query data. The three search MT models all use the same generic MT system which is trained on the generic data, then they are fine-tuned on different set of domain-specific data for domain adaptation: M_0 is the baseline model which is fine-tuned on the query data. M_1 is fine-tuned on the query data and query data with prefix context. M_2 is fine-tuned on the query data and query data with selected prefix context.

5. Results and Analysis

MT Metrics: Table 2 shows the MT metrics[6]. For engb-dede, we observe both models M_1 and M_2 have better query translation quality than the baseline model M_0 with +0.7 and +0.4 BLEU scores respectively as well as COMET and chrF. For the other two language pairs, models M_1 and M_2 do not have higher MT metrics than baseline. Based on the Levenshtein distance as shown in table 1 for the query session data analysis, there are likely much more contextual queries that are spelling variant queries instead of content queries for the frca-enca and

[5] All of the operation (insertion, deletion and substitution) costs are set to 1

[6] All the MT metrics are computed with lowercase

Lang pair	Model	BLEU	chrF	COMET
engb-dede	M_0 (baseline)	55.1	82.9	86.7
	M_1 (full context)	**55.7**	**83.0**	**87.1**
	M_2 (selected context)	55.5	83.0	86.9
frca-enca	M_0 (baseline)	**54.2**	**80.9**	**87.6**
	M_1 (full context)	53.1	80.4	86.5
	M_2 (selected context)	53.8	80.6	87.3
ptbr-enus	M_0 (baseline)	**53.2**	**78.8**	**85.5**
	M_1 (full context)	53.0	78.6	84.4
	M_2 (selected context)	53.0	78.8	84.6

Table 2
Translation quality metric results of the MT models.

ptbr-enus. For frca-enca, model M_2 has improved +0.7 BLEU than Model M_1 after we filter 19% data with lower Levenshtein edit distance than the mean of the data set. Therefore, it signals that spelling variants are not beneficial to the MT training, which is consistent with our initial hypothesis. For ptbr-enus, we also observe the same pattern as the frca-enus. In addition, the query data with prefix context is also much smaller than the other two language pairs so the overall influence on the overall MT metrics is much smaller accordingly.

Search Metrics: Table 4 shows the search metrics[7] for the translated queries from the three models. Overall, the search metrics are consistent with the MT metrics across the language pairs except for ptbr-enus. For ptbr-enus, model M_1 has higher MAP, Precision, Recall and F1 although it has slightly lower MT metrics than the baseline Model M_0. Therefore, the query translation from the model M_1 is slightly more suitable for the down stream search tasks. We also observe the scale of improvement in search metrics is smaller than the MT metrics because search ecosystems have different tolerance of the query translation quality across different language pairs, which is also analyzed in a previous study [27].

Improved Query Translations: We further investigate the improved query translations from the model M_1 for the engb-dede language pair. We observe cases which have shown the the model M_1 can learn the word sense better with the context during training. As table 3 shows, for the English query *twist off glasses*, the word *glasses* can be either the glasses for eyes or the glass containers. In this query, it refers to the glass container that can be twisted open. The model M_1 trained with context can translate the query into the correct word sense whereas the baseline model M_0 translates it into eye glasses. Another example is *face mask christmas*, which refers to a mask with Christmas patterns. While model M_0 translates it into the cosmetic face mask, model M_1 translates the query to the correct product. We also observe a number of cases where the model M_1 can translate with better word choice and forms. For example, the translation of the query *hand cream dispenser* is translated as *handcreme spender* whereas it is translated as *handsahnespender* by the baseline model. In this case, the term *handcreme* is more

[7] All the search metrics are computed using Top-16 search results

Query Source (English)	Query Translation (German)	
	Model M_0	**Model** M_1
twist-off glasses	twist-off brille	twist-off gläser
hand cream dispenser	handsahnespender	handcreme spender
face mask christmas	gesichtsmaske weihnachten	mundschutz weihnachten
car wash	autowaschanlage	autowäsche
sanding block	schleifbock	schleifklotz
wastepaper bin	papiertonne	papierkorb
flower box holder	blumenkastenhalter	blumenkastenhalterung
storage tins kitchen	aufbewahrungsdosen küche	vorratsdosen küche
fidget toy	fidget toy	zappeln spielzeug
dental mirror	dental spiegel	zahnspiegel

Table 3
Query translation comparison

Lang pair	Model	nDCG	MAP	Precision	Recall	F1
engb-dede	M_0	54.98	46.61	29.43	2.76	4.94
	M_1	**55.20**	**46.71**	**29.48**	**2.77**	**4.96**
	M_2	54.92	46.54	29.43	2.76	4.95
frca-enca	M_0	**56.41**	**47.13**	**29.79**	**3.34**	**5.88**
	M_1	55.97	46.84	29.57	3.30	5.82
	M_2	55.98	46.81	29.68	3.32	5.85
ptbr-enus	M_0	61.19	53.86	34.76	3.11	5.52
	M_1	61.16	**53.93**	**34.81**	**3.12**	**5.53**
	M_2	60.96	53.90	34.64	3.09	5.49

Table 4
Search metric results of the MT models. M_0: Baseline model, M_1 model with prefix context, M_2 model with selected prefix context.

proper word choice than *handsahne*. Furthermore, there are also cases where the model M_1 can return correct translations whereas the baseline model struggles to translate. For example, queries *fidget toy* and *dental mirror* remain untranslated or partially untranslated while model M_0 and M_1 returns a full translation.

6. Related work

Previous studies have shown the benefits of using extended source language context as well as bilingual context extensions in attention-based neural machine translation[18]. Other studies have also shown extended and modified deep network-based MT models can take advantage of document-level context [28, 29, 28, 30, 31, 32, 33, 34]. In the search community, leveraging queries from search session to improve query understanding tasks such as query rewriting, query reformulation and search intent detection is a well-studied field of research [35, 36, 37, 38]. However, there has not been extensive study on utilizing contextual information to improve machine translation for the purpose of improved search retrieval. Our contribution is that by employing session-based queries to improve machine translation quality, we are able to improve downstream search retrieval tasks.

7. Conclusion and future work

In this paper, we first propose two main categories of contextual queries based on our data exploration and analysis, then propose to use queries from search sessions as contextual signal to improve MT for e-commerce multilingual search. Our results show that using content variant contextual queries to augment training data as contextual signals can be beneficial to training MT for query translation, while using more spelling variant contextual queries can have negative impact. While as a pilot study our experiment is limited to three language pairs, we will further explore the potential of employing contextual signals for a wider range of language pairs in future work.

References

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

[2] B. Zhang, Improve MT for search with selected translation memory using search signals, in: Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track), Association for Machine Translation in the Americas, Orlando, USA, 2022, pp. 123–131. URL: https://aclanthology.org/2022.amta-upg.9.

[3] K. Hui, A. Yates, K. Berberich, G. de Melo, PACRR: A position-aware neural IR model for relevance matching, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1049–1058. URL: https://aclanthology.org/D17-1110. doi:10.18653/v1/D17-1110.

[4] R. McDonald, G. Brokos, I. Androutsopoulos, Deep relevance ranking using enhanced document-query interactions, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1849–1860. URL: https://aclanthology.org/D18-1211. doi:10.18653/v1/D18-1211.

[5] P. Nigam, Y. Song, V. Mohan, V. Lakshman, W. A. Ding, A. Shingavi, C. H. Teo, H. Gu, B. Yin, Semantic product search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2876–2885. URL: https://doi.org/10.1145/3292500.3330759. doi:10.1145/3292500.3330759.

[6] H. Lu, Y. Hu, T. Zhao, T. Wu, Y. Song, B. Yin, Graph-based multilingual product retrieval in E-commerce search, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry

Papers, Association for Computational Linguistics, Online, 2021, pp. 146–153. URL: https://aclanthology.org/2021.naacl-industry.19. doi:10.18653/v1/2021.naacl-industry.19.

[7] S. Li, F. Lv, T. Jin, G. Lin, K. Yang, X. Zeng, X.-M. Wu, Q. Ma, Embedding-based product retrieval in taobao search, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3181–3189.

[8] J.-Y. Nie, Cross-language information retrieval, Synthesis Lectures on Human Language Technologies 3 (2010) 1–125.

[9] A. Rücklé, K. Swarnkar, I. Gurevych, Improved cross-lingual question retrieval for community question answering, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 3179–3186. URL: https://doi.org/10.1145/3308558.3313502. doi:10.1145/3308558.3313502.

[10] S. Saleh, P. Pecina, Document translation vs. query translation for cross-lingual information retrieval in the medical domain, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6849–6860. URL: https://aclanthology.org/2020.acl-main.613. doi:10.18653/v1/2020.acl-main.613.

[11] T. Bi, L. Yao, B. Yang, H. Zhang, W. Luo, B. Chen, Constraint translation candidates: A bridge between neural query translation and cross-lingual information retrieval, 2020. arXiv:2010.13658.

[12] Z. Jiang, A. El-Jaroudi, W. Hartmann, D. Karakos, L. Zhao, Cross-lingual information retrieval with BERT, in: Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020), European Language Resources Association, Marseille, France, 2020, pp. 26–31. URL: https://aclanthology.org/2020.clssts-1.5.

[13] H. Zhang, L. Tan, Textual representations for crosslingual information retrieval, in: Proceedings of the 4th Workshop on e-Commerce and NLP, Association for Computational Linguistics, Online, 2021, pp. 116–122. URL: https://aclanthology.org/2021.ecnlp-1.14. doi:10.18653/v1/2021.ecnlp-1.14.

[14] Y. Lv, C. Zhai, Adaptive relevance feedback in information retrieval, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 255–264.

[15] O. Vechtomova, Y. Wang, A study of the effect of term proximity on query expansion, Journal of Information Science 32 (2006) 324–333.

[16] A. Goldfarb, D. Trefler, et al., Artificial intelligence and international trade, The economics of artificial intelligence: an agenda (2019) 463–492.

[17] E. Brynjolfsson, X. Hui, M. Liu, Does machine translation affect international trade? evidence from a large digital platform, Management Science 65 (2019) 5449–5460.

[18] J. Tiedemann, Y. Scherrer, Neural machine translation with extended context, arXiv preprint arXiv:1708.05943 (2017).

[19] V. I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet physics doklady, volume 10, Soviet Union, 1966, pp. 707–710.

[20] H. Zhang, L. Tan, A. Misra, Evaluating machine translation in cross-lingual E-commerce search, in: Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Orlando, USA, 2022, pp. 322–334. URL: https://aclanthology.org/2022.amta-research.25.

[21] L. Wu, D. Hu, L. Hong, H. Liu, Turning clicks into purchases: Revenue optimization for product search in e-commerce, SIGIR '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 365–374. URL: https://doi.org/10.1145/3209978.3209993. doi:10.1145/3209978.3209993.

[22] M. Post, A call for clarity in reporting BLEU scores, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 186–191. URL: https://aclanthology.org/W18-6319. doi:10.18653/v1/W18-6319.

[23] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213. doi:10.18653/v1/2020.emnlp-main.213.

[24] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049. doi:10.18653/v1/W15-3049.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[26] T. Domhan, M. Denkowski, D. Vilar, X. Niu, F. Hieber, K. Heafield, The sockeye 2 neural machine translation toolkit at AMTA 2020, in: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), Association for Machine Translation in the Americas, Virtual, 2020, pp. 110–115. URL: https://aclanthology.org/2020.amta-research.10.

[27] B. Zhang, A. Misra, Machine translation impact in E-commerce multilingual search, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track, Association for Computational Linguistics, Abu Dhabi, UAE, 2022, pp. 99–109. URL: https://aclanthology.org/2022.emnlp-industry.8.

[28] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, Y. Liu, Improving the transformer translation model with document-level context, arXiv preprint arXiv:1810.03581 (2018).

[29] L. Wang, Z. Tu, A. Way, Q. Liu, Exploiting cross-sentence context for neural machine translation, arXiv preprint arXiv:1704.04347 (2017).

[30] P. Fernandes, K. Yin, G. Neubig, A. F. Martins, Measuring and increasing context usage in context-aware machine translation, arXiv preprint arXiv:2105.03482 (2021).

[31] E. Voita, R. Sennrich, I. Titov, When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion, arXiv preprint arXiv:1905.05979 (2019).

[32] Z. Tu, Y. Liu, S. Shi, T. Zhang, Learning to remember translation history with a continuous cache, Transactions of the Association for Computational Linguistics 6 (2018) 407–420.

[33] S. Maruf, A. F. Martins, G. Haffari, Selective attention for context-aware neural machine translation, arXiv preprint arXiv:1903.08788 (2019).

[34] S. Maruf, G. Haffari, Document context neural machine translation with memory networks, arXiv preprint arXiv:1711.03688 (2017).

[35] S. Zuo, Q. Yin, H. Jiang, S. Xi, B. Yin, C. Zhang, T. Zhao, Context-aware query rewriting for improving users' search experience on e-commerce websites, 2022. arXiv:2209.07584.

[36] B. Mitra, Exploring session context using distributed representations of queries and reformulations, in: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 3–12.

[37] S. Chawla, P. Bedi, Improving information retrieval precision by finding related queries with similar information need using information scent, in: 2008 First International Conference on Emerging Trends in Engineering and Technology, IEEE, 2008, pp. 486–491.

[38] F. Baskaya, H. Keskustalo, K. Järvelin, Modeling behavioral factors ininteractive information retrieval, in: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 2013, pp. 2297–2302.