Machine Translation Impact in E-commerce Multilingual Search

Bryan Hang Zhang
Amazon.com
bryzhang@amazon.com

Amita Misra
Amazon.com
misrami@amazon.com

Abstract

Previous work suggests that performance of cross-lingual information retrieval correlates highly with the quality of Machine Translation. However, there may be a threshold beyond which improving query translation quality yields little or no benefit to further improve the retrieval performance. This threshold may depend upon multiple factors including the source and target languages, the existing MT system quality and the search pipeline. In order to identify the benefit of improving an MT system for a given search pipeline, we investigate the sensitivity of retrieval quality to the presence of different levels of MT quality using experimental datasets collected from actual traffic. We systematically improve the performance of our MT systems quality on language pairs as measured by MT evaluation metrics including Bleu and Chrf to determine their impact on search precision metrics and extract signals that help to guide the improvement strategies. Using this information we develop techniques to compare query translations for multiple language pairs and identify the most promising language pairs to invest and improve.

1 Introduction

Multilingual search capability is essential for modern e-commerce product discovery (Lowndes and Vasudevan, 2021; Zhang, 2022). Localization of e-commerce sites have led users to expect search engines to handle multilingual queries. Recent proposals of cross-lingual information retrieval handle multilingual queries, and language-agnostic cross-borders product indexing has gained traction with neural search engines (Hui et al., 2017; McDonald et al., 2018; Nigam et al., 2019a; Lu et al., 2021; Li et al., 2021), but legacy e-commerce search indices are still built on monolingual product information and support for multilingual search is bridged using Query translation (Nie, 2010; Rücklé et al., 2019; Saleh and Pecina, 2020; Bi et al., 2020; Jiang et al., 2020; Zhang and Tan, 2021).

Query translation allows users to look up information represented in documents written in a languages different from the language of the query. It takes as input the query typed in source or query language and returns a translated query to the search engine to retrieve documents in the target language. It follows that query translation plays a key role and its output significantly affects the retrieval results.

Previous studies have demonstrated performance of CLIR (Cross-Lingual Information Retrieval) correlates highly with the quality of the Machine Translation (MT), and improving the quality of MT improves retrieval quality (Goldfarb et al., 2019; Brynjolfsson et al., 2019). However, these evaluations are done separately for each task. This leaves a large gap in understanding the impact of improving MT quality iteratively on CLIR performance in a real time industrial setting. Since machine translation is used here as interim application, the objectives of the retrieval task may have varying levels of tolerance to the inherent translation quality. Information retrieval evaluation usually involves human-annotated relevance labels of search results candidates. In an industry setting, annotating a representative sample is a time consuming and expensive task, particularly during iterative improvement of MT for the search use case. Additionally, a general-purpose MT evaluation metric may not necessarily adapt to the query evaluation for downstream retrieval task.

To address these above concerns, we propose an MT evaluation framework to build an e-commerce specific CLIR test set. It exploits behavioural signals from search retrieval results to evaluate MT quality for a given query. In order to identify the benefit of improving an MT system, we further investigate the sensitivity of retrieval quality to the presence of different levels of MT quality as measured by Bleu, and Chrf using experimental datasets collected from actual traffic. Based on

these experiments, we recommend the pairs that are worth continued investment in improving MT systems for search. Our main contributions are:

- A **rank-based evaluation framework** to evaluate MT in CLIR through ranking-based search metrics using behavioral signals (from the store of the target language) as a proxy to relevance information without any human annotation; this framework can be used to create e-commerce CLIR test set at scale.
- A **method to measure the MT launching impact** on the e-commerce CLIR ecosystems for a given language pair. This can be used to identify and prioritize the high impact language pairs for more investment in the MT improvement.
- A **method to measure the MT improvement impact** on the e-commerce CLIR ecosystems for a given language pair. It signals the strategy to be used for MT improvement, either a comprehensive strategy focusing on the overall query traffic or a specific one targeting a smaller percentage of query traffic or a combination of both strategies.

This paper is organized as following: we propose a rank-based evaluation framework in section 2. We propose two MT impact rates, MT launching impact rate and MT improvement MT rate respectively in section 3. Section 4 is the experiment with 12 language pairs from 6 stores. Section 5 is the results and analysis. We defer related work to Section 6 where we compare it with our proposed work. We draw a conclusion in Section 7.

2 Cross-Lingual Information Retrieval (CLIR) Evaluation Framework for E-commerce Product Search

Different from static test sets in academia, industrial search applications are dynamic as user queries and behavioral signals change with world trends. Moreover, product inventory is dynamic, changes often and quickly.

A previous study (Sloto et al., 2018) proposes the traditional Normalized Discounted Cumulative Gain (nDCG) for CLIR using all search results from the reference translation as relevance ground truth to compute nDCG for MT translation (aka nDCG-MT). However, their approach imposes a strong assumption that the top-k search results from reference translation are all relevant to the query and relevance is inversely scaled by the ranking of the results.

Although behavioral signals from users' clicks and purchases are useful proxy (Wu et al., 2018) to expensive human relevance annotations, these are dynamic and change according to the product life cycle and seasonal business trends. These behavioral signals need to be updated at regular cadence to accurately represent relevance information needed to compute search metrics.

We introduce a ranking-based evaluation framework through search ranking metrics using behavioral signals as a proxy to relevance information without any human annotation; To the best of our knowledge, there is no systematic study on cross-lingual information retrieval for e-commerce search that neither requires ground-truth click/purchase information nor human annotated relevance data.

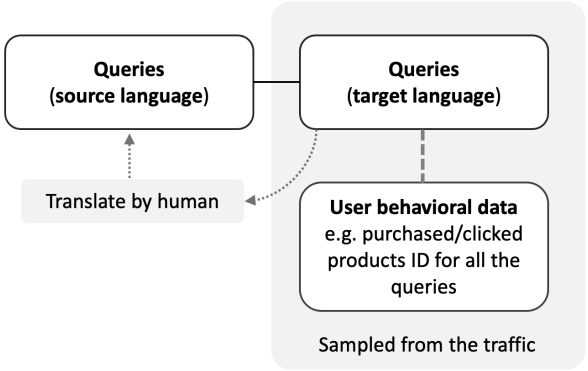

Figure 1: Test set creation workflow

Figure 1 illustrates the test sets creation workflow for MT evaluation in E-commerce CLIR:

1. Create a sample of query data from the historical search traffic in the target language (the language that the search index is built on). Empirically, we recommend to sample that queries from the top 30%, bottom 30% and the middle 40% in frequency bins to better simulate the user traffic. *We refer to these queries as* Q_{ref}.
2. To allow computation of traditional relevance metrics, aggregate the clicks and/or purchase product IDs associated with the queries, if they are available. We refer to the products IDs associated with the query and their click/purchase frequency as P_{id} and P_{freq}.
3. Create human reference translation of the

search queries sample in the source language (the language that users will be searching in). *We refer to these human translated queries as* Q_{src}.

We propose the following evaluation framework with the test sets created above to evaluate machine translation in the context of CLIR for e-commerce queries.

1. Translate the Q_{src} with the MT model in consideration. *We refer to these machine translated queries as* Q_{mt}.
2. Search for the candidate products using the machine translated queries Q_{mt}; retrieving top-k search result R_{mt}
3. Use P_{id} and P_{freq} (as ground truth) with R_{mt} to compute traditional relevance based metrics such as nDCG.

3 MT impact in the search ecosystem

3.1 The range of MT impact on search

As mentioned, the downstream search pipeline consists of a large number of components, which altogether has different levels of tolerance for query translation quality. Hence, it is important to estimate the range of query translation impact on the search ecosystem in consideration. With the test sets from the creation workflow in Section 2, we propose to measure the rank-based search metrics such as nDCG of source queries Q_{src} as the lower bound of the MT impact, which serves a baseline for the impact of MT translated query on search, and measure the search metrics of human reference query translations Q_{ref} as the upper bound.

3.2 MT launching impact measurement

We expect that launching an MT system in a search ecosystem of different language pairs can have different levels of positive impact on the search result quality. Therefore, given a language pair (*e.g. enus-jajp*), we propose the **MT launching impact rate** to quantify the MT impact on the search ecosystem (*e.g. jp*). MT launching impact rate (I_{MT}) is defined as in Equation 1.

$$I_{MT} = \frac{\Delta S}{\Delta T} \tag{1}$$

$$\Delta S = S_{adapt} - S_{source} \tag{2}$$

$$\Delta T = T_{adapt} - T_{source} \tag{3}$$

where, ΔS is the search result improvement from source queries S_{source} to the query translation from a fine-tuned MT S_{adapt} (as Equation 2), S_{source} and S_{adapt} can be common search metrics such as nDCG, ΔT is the respective translation quality improvement from the source queries T_{source} to the fine-tuned MT query translations T_{adapt} (as equation 3), T_{source} and T_{adapt} can be MT evaluation metrics such as Bleu or Chrf.

We propose the following three groups for language pairs based on their MT impact rate:

- **High-impact language pairs**: Search ecosystems of high-impact language pairs are less tolerant to languages different from the search index language, and more sensitive to the query translation quality. Launching or improving an MT system of those language pairs in the respective search pipeline is more likely to improve the search results.
- **Medium-impact language pairs**: Search ecosystems of medium-impact language pairs are somewhat sensitive to the query translation quality, though not as much as high-impact language pairs.
- **Low-impact language pairs**: Search ecosystems of low-impact language pairs are more robust to different languages and translation quality, and the presence of an MT in the search pipeline has less or little impact on the search result improvement.

3.3 MT improvement impact

We experimented with two improvement strategies for MT in the e-commerce CLIR product search: one is **comprehensive improvement (CI)**, the other is **specific improvement (SI)**. CI usually focuses on the overall improvement in translation quality and targets the entire query traffic. The CI strategies usually involve a change of model architecture or training techniques, etc; SI usually focuses on the improvement of the specific aspects of the query translation quality, and targets a fraction of query traffic. The SI strategies are not necessarily language-agnostic, for example, it can be solving a smaller transliteration problem in a given language, or a brand term preservation improvement for a given language pair.

We propose **The MT improvement impact rate** to quantify the impact of MT comprehensive improvement ($I_{improve}$) on search improvement as in

Equation 4, which can provide signals to choose the right MT improvement strategy for a given language pair.

$$I_{improve} = \frac{\Delta S'}{\Delta T'} \quad (4)$$

$$\Delta S' = S_{adapt} - S_{generic} \quad (5)$$

$$\Delta T' = T_{adapt} - T_{generic} \quad (6)$$

where, $\Delta S'$ is the search result improvement from generic MT query translations $S_{generic}$ to the fine-tuned MT query translations S_{adapt} (as in Equation 5), $S_{generic}$ and S_{adapt} can be the common search metrics such as nDCG; $\Delta T'$ is the respective translation quality improvement from generic MT query translations $T_{generic}$ to the fine-tuned MT query translations T_{adapt} (as in Equation 6), $T_{generic}$ and T_{adapt} can be MT evaluation metrics such as Bleu or Chrf.

Language pairs with higher improvement rate signals both the CI and SI of MT are likely to have positive impact on search. Those with lower rate may benefit more from the focusing on SI for a targeted group of queries from the traffic.

4 Experiment

Language pairs and locales: We selected 12 language pairs from 6 stores for our experiments as seen in Table 1.

Lang pair	Store	Lang pair	Store
esmx-enus	US	ptpt-eses	Spain
ptbr-enus	US	frca-enca	Canada
kokr-enus	US	nlnl-dede	Germany
dede-enus	US	trtr-dede	Germany
mlin-enin	India	engb-dede	Germany
knin-enin	India	enus-jajp	Japan

Table 1: Selected 12 language pairs from 6 stores

Test data: The test data is created as described in Section 2. The test set comprises 4000 queries (as reference query translation) per store (e.g. enus), each query is translated into their respective language pairs (e.g. enus -> kokr, enus -> dede). We have also stored the purchased product IDs associated with the queries of the store (e.g. US). We use sampled purchased product ID associated with reference queries as relevant product, and the logarithm of the frequencies of purchased product as the relevance score.

Machine Translation (MT) models: We trained two models per language pair: (i) a *generic MT* system trained on general news and internal crawled data with (ii) a *domain-specific MT* that is fined tuned on human translated search queries and synthetically generated query translations through back-translation. These in-house MT models are trained on proprietary data using vanilla transformer architecture (Vaswani et al., 2017) with Sockeye MT toolkit (Domhan et al., 2020).[1]

Metric hyper-parameters: We set K to 16 for the top-k search results, using the top-16 products in the search results to compute nDCG@16.

MT metrics: Tables 3 and 4 in the appendix present the MT quality metrics Bleu[2] and Chrf; Table 5 in the appendix presents search performance metric normalized nDCG@16.[3].

MT launching and improvement impact rates: With aforementioned metrics, the lower and higher bounds of nDCG@16 of MT impact are presented in Table 6. MT launching impact and improvement rates are computed using nDCG@16 with and Chrf respectively, as in Table 2 in the appendix.

	MT launching impact		MT improvement impact	
Language pair	**ΔnDCG/ ΔBleu**	**ΔnDCG/ ΔChrf**	**ΔnDCG/ ΔBleu**	**ΔnDCG/ ΔChrf**
ptpt-eses	0.11	0.15	0.19	0.70
enus-jajp	0.25	0.18	0.78	1.09
engb-dede	0.29	0.32	0.09	0.13
frca-enca	0.31	0.23	0.35	0.60
nlnl-dede	0.47	0.43	0.32	0.69
esmx-enus	0.50	0.34	0.34	0.64
ptbr-enus	0.62	0.56	0.28	1.01
dede-enus	0.62	0.66	0.33	0.61
knin-enin	0.72	0.59	0.19	0.60
trtr-dede	0.85	0.43	0.24	0.43
kokr-enus	0.98	0.49	0.33	0.39
mlin-enin	1.04	0.59	0.74	0.72

Table 2: MT launching impact and improvement impact rates

5 Results and Analysis

For the MT launching impact, we rank the language pairs in the descending order according to the MT launching impact rate as well as the impact range respectively, as in Table 7 in the appendix. We observe Bleu and Chrf can give a similar ranking

[1]For the purpose of this paper, we are less concerned with the accuracy of the MT models and more interested in the difference in the MT quality as per measured by traditional MT metrics and their evaluation based on our proposed framework. Thus the brevity in the model description.

[2]SacreBleu version 2.0.0 (Post, 2018)

[3]Both the nDCG@16 and Chrf are scaled to 0-100 for the computation convenience

with small difference, so the following analysis is based on the MT launching impact from Δ nDCG/ ΔBleu for simplicity. For the MT improvement impact rate, we observe that Bleu makes value scale smaller than Chrf. We will use ΔnDCG/ ΔBleu for the following analysis.

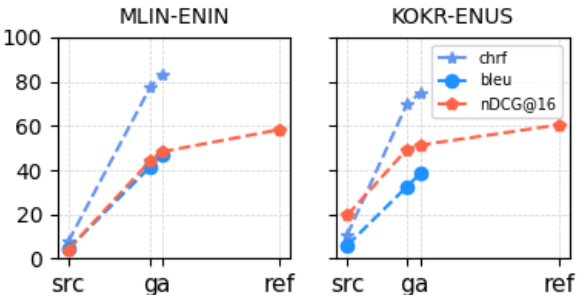

Figure 2: Language pairs where the MT have bigger impact on search pipeline

We observe that MT of language pairs such as *mlin-enin, kokr-enus* have higher launching impact rate and should be labeled as the high-impact language pairs. For *mlin-enin*, the MT launching impact rate is 1.04, which signals one point Bleu increase in translation quality can gain slightly more than one point of search improvement. Figure 2 (In Figure 2, 3, 4 , "src" refers to source query, "g" refers to generic mt, "a" refers to the adapted (fine-tuned) MT, "ref" refers to the human translation. The axis is scaled according to the Bleu score from 0-100.) illustrates the higher impact language pairs, the range of the MT impact is much bigger, search ecosystems are very responsive to the presence of MT system in the search pipeline, MT and search metrics have similar trending. *mlin-enin* has a much higher improvement rate of 0.74, the ecosystem of the search of this language pair can potentially benefit from both comprehensive improvement (CI) and specific improvement (SI) in the MT. Meanwhile, *kokr-enus* has a much lower improvement rate of 0.33, which signals this search is more likely to benefit from SI than CI.

Language pairs such as *nlnl-dede, frca-enca* should be considered as the decent impact language pairs. As illustrated in Figure 3, both have smaller MT impact range and the launching impact rates are high but not quite as the high impact language pairs. As Bleu and Chrf increase from source query to generic MT to fine-tuned MT, nDCG@16 increases slower. Both language pairs have relative lower improvement impact rate which is around 0.3, that signals search of these two language

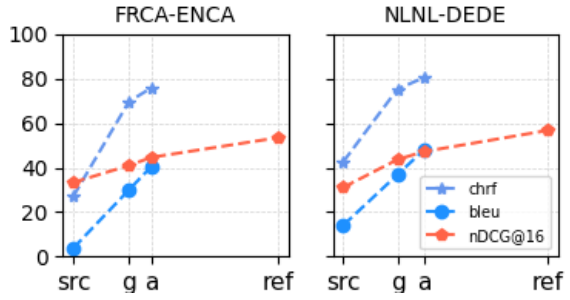

Figure 3: Language pairs where the MT have decent impact on search pipeline

pairs are more likely to benefit from SI than CI. Language pairs such as *ptpt-eses, enus-jajp* should

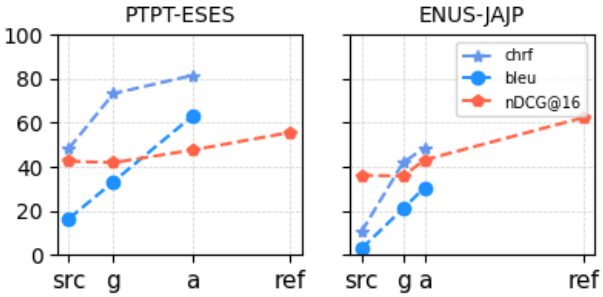

Figure 4: Language pairs where the MT have lower impact on search pipeline

be labeled lower impact language pairs based on the lower launching impact rate. For *ptpt-eses*, one point Bleu increase in translation quality can only achieve 0.11 point of search improvement. Both have smaller launching impact range, thus, search ecosystems are not very responsive to the MT quality improvement. As Bleu and Chrf increase from source query to generic MT to fine-tuned MT, nDCG@16 increases much slower and the trend line is almost flat as Figure 4. In principle, low-impact language arcs might not be prioritized for MT improvement. If there is a need to improve those MT for search, *ptpt-eses* has a much lower MT impact rate of 0.19, so search is likely to benefit from the SI for the MT, whereas *enus-jajp* has much higher improvement rate of 0.78, the search may still benefit from CI as well as SI. Figure 5 and 6 in the appendix are the plots for all other language pairs.

A/B testing: We have also conducted parallel online A/B testing for the following language pairs: *enus-jajp, ptpt-eses, frca-enca, mlin-enin, nlnl-dede, engb-dede*. For each language pair, we have deployed two fine-tuned MT systems and

integrated them into the search pipeline for the designated store, and the MT system with the comprehensive improvement has higher off-line MT metrics (+5 Bleu points on average) than the baseline model. The A/B testing lasted for 4 weeks on average for all the experiments. For the high impact language pairs, the improved MT systems have seen large increases in business metrics, such as, Order Product Sales (OPS), composite contribution profit (CCP), compared to the baseline model, and have much larger positive impact on the search result quality. For the low impact language pairs, we observe much smaller or even no impact at all. Overall, the A/B testing results are consistent with the MT launching impact rate results we have computed. Moreover, for ptpt-eses and nlnl-dede, we also conducted another round A/B testing with the same experiment setup except using MT with specific improvement to compare with the baseline models. Those two improved MT enhanced the terminology translation of 3-5% of query traffic. The results are consistent with our hypothesis that the MT with SI improvement has much more impact than the MT with CI improvement.

6 Related Work

Machine Translation is necessary to bridge the gap between query translation and cross-lingual information retrieval (Bi et al., 2020). Query translation a key component in large e-commerce stores, previous studies have demonstrated that better translation quality improves retrieval accuracy (Goldfarb et al., 2019; Brynjolfsson et al., 2019).

Queries are naturally short and search engines usually have preferred word choices and collocations based on users' query patterns (Lv and Zhai, 2009; Vechtomova and Wang, 2006). This complicates the evaluation of machine translation for cross-lingual information retrieval in the context of 'fitting in well to the search index'. While machine translation evaluation is well-studied, translation evaluation in downstream task requires more attention especially in the e-commerce cross-lingual information retrieval.

Traditionally, information retrieval evaluation relies on behavioral signals as ground truth to measure relevance of search results; mean reciprocal ranking (MRR), mean average precision (MAP), normalized discounted cumulative gain (nDCG) (Järvelin and Kekäläinen, 2002; Wu et al., 2018; Nigam et al., 2019b).

Previous studies in cross-lingual information retrieval (CLIR) evaluation relies on pre-annotated datasets that are relatively small and specific to domains outside of e-commerce; for example, the CLEF eHealth test sets (Saleh and Pecina, 2018; Suominen et al., 2018; Zhang et al., 2013) and Wikipedia cross-lingual test set (Sas et al., 2020). Although Sloto et al. (2018) proposed the nDCG-MT metric that leveraged on the reference translation to measure search results relevance, reliance on the ground truth data is still necessary. In pursuit of a more effective approach, we integrate CLIR and MT more closely and evaluate them in an end-to-end task. Our proposed method allows us to fully-automate the evaluation and study the impact of improving MT on CLIR by collecting organic queries in the target language of the e-commerce service and use reference results of these queries as a proxy to human annotation.

7 Conclusion

In this paper, we propose an evaluation framework for MT in the E-commerce multilingual product search through ranking-based search metrics using behavioral signals as proxy relevance information without any human notation, which makes it practical to iteratively improve MT models for the search use case and evaluate them frequently off-line. This framework can also be used to create cross-lingual information retrieval (CLIR) test sets for e-commerce at scale. We also propose a method to measure off-line the MT launching impact and and improvement impact rate on search. The former can identify the the high-impact language pairs can be prioritized with more investment in the MT improvement. These experiments can help select the most promising improvement strategy either comprehensive or specific improvement or combination of both to bring a larger impact on the search performance of a given language pair. We have experimented with the proposed evaluation framework and MT impact measuring method on 12 language pairs from 6 stores, and identified the high language pairs of different impact on search and assigned potential improvement strategies. The results are consistent with on-line A/B testing.

References

Tianchi Bi, Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, and Boxing Chen. 2020. Constraint

translation candidates: A bridge between neural query translation and cross-lingual information retrieval.

Erik Brynjolfsson, Xiang Hui, and Meng Liu. 2019. Does machine translation affect international trade? evidence from a large digital platform. *Management Science*, 65(12):5449–5460.

Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. 2020. 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)*, pages 110–115, Virtual. Association for Machine Translation in the Americas.

Avi Goldfarb, Daniel Trefler, et al. 2019. Artificial intelligence and international trade. *The economics of artificial intelligence: an agenda*, pages 463–492.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1049–1058, Copenhagen, Denmark. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos, and Lingjun Zhao. 2020. Cross-lingual information retrieval with BERT. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 26–31, Marseille, France. European Language Resources Association.

Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3181–3189.

Mike Lowndes and Aditya Vasudevan. 2021. Market guide for digital commerce search.

Hanqing Lu, Youna Hu, Tong Zhao, Tony Wu, Yiwei Song, and Bing Yin. 2021. 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*, pages 146–153, Online. Association for Computational Linguistics.

Yuanhua Lv and ChengXiang Zhai. 2009. Adaptive relevance feedback in information retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 255–264.

Ryan McDonald, George Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1849–1860, Brussels, Belgium. Association for Computational Linguistics.

Jian-Yun Nie. 2010. Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1):1–125.

Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian (Allen) Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019a. Semantic product search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '19, page 2876–2885, New York, NY, USA. Association for Computing Machinery.

Priyanka Nigam, Yiwei Song, Vijai Mohan, Vihan Lakshman, Weitian, Ding, Ankit Shingavi, Choon Hui Teo, Hao Gu, and Bing Yin. 2019b. Semantic product search.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Andreas Rücklé, Krishnkant Swarnkar, and Iryna Gurevych. 2019. Improved cross-lingual question retrieval for community question answering. In *The World Wide Web Conference*, WWW '19, page 3179–3186, New York, NY, USA. Association for Computing Machinery.

Shadi Saleh and Pavel Pecina. 2018. Cuni team: Clef ehealth consumer health search task 2018. In *CLEF (Working Notes)*.

Shadi Saleh and Pavel Pecina. 2020. 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*, pages 6849–6860, Online. Association for Computational Linguistics.

Cezar Sas, Meriem Beloucif, and Anders Søgaard. 2020. WikiBank: Using Wikidata to improve multilingual frame-semantic parsing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4183–4189, Marseille, France. European Language Resources Association.

Steve Sloto, Ann Clifton, Greg Hanneman, Patrick Porter, Donna Gates, Almut Silja Hildebrand, and Anish Kumar. 2018. Leveraging data resources for cross-linguistic information retrieval using statistical machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 223–233.

Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. 2018. Overview of the clef ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Olga Vechtomova and Ying Wang. 2006. A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333.

Liang Wu, Diane Hu, Liangjie Hong, and Huan Liu. 2018. Turning clicks into purchases: Revenue optimization for product search in e-commerce. SIGIR '18, page 365–374, New York, NY, USA. Association for Computing Machinery.

Bryan Zhang. 2022. 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)*, pages 123–131, Orlando, USA. Association for Machine Translation in the Americas.

Hang Zhang and Liling Tan. 2021. Textual representations for crosslingual information retrieval. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 116–122, Online. Association for Computational Linguistics.

Lei Zhang, Achim Rettinger, Michael Färber, and Marko Tadić. 2013. A comparative evaluation of cross-lingual text annotation techniques. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 124–135. Springer.

A Appendix

sacreBleu			
Language pair	**source**	**generic MT**	**adapted MT**
trtr-dede	6.4	23.4	28.8
enus-jajp	2.8	21.1	30.6
esmx-enus	2.6	26.6	33.3
kokr-enus	6.02	32.53	38.39
frca-enca	3.77	30.01	40.46
ptbr-enus	3.7	26.8	41.91
mlin-enin	4.41	41.7	47.02
nlnl-dede	14.09	36.87	48.11
dede-enus	6.88	46.74	60.93
ptpt-eses	16.49	33.28	63.08
engb-dede	10.1	45.61	63.08
knin-enin	2.77	52.02	71.27

Table 3: MT metric - Bleu for source queries and query MT translations

Chrf			
language pair	**source**	**generic mt**	**adapted mt**
dede-enus	30.49	73.82	81.36
engb-dede	33.08	69.68	80.99
enus-jajp	10.49	41.91	48.67
esmx-enus	24.92	65.62	69.19
frca-enca	27.04	69.72	75.85
kokr-enus	10.7	69.79	74.75
mlin-enin	7.64	77.7	83.19
nlnl-dede	42.63	75.34	80.58
ptbr-enus	25.66	64.2	68.33
ptpt-eses	48.37	73.26	81.52
trtr-dede	23.08	64.27	67.3
knin-enin	5.29	82.67	88.62

Table 4: MT metric -Chrf for source queries and query MT translations

nDCG@16				
Language pair	**source**	**generic MT**	**adapted MT**	**ref**
enus-jajp	36.2	35.80	43.19	62.30
frca-enca	33.34	40.98	44.64	53.47
trtr-dede	26.8	44.60	45.90	63.90
nlnl-dede	31.11	43.67	47.26	56.76
ptpt-eses	42.53	41.89	47.64	55.65
mlin-enin	4.00	44.38	48.34	58.28
ptbr-enus	27.2	46.71	50.89	60.28
kokr-enus	19.59	49.38	51.29	60.42
dede-enus	17.78	46.91	51.54	60.27
knin-enin	2.90	48.7	52.27	58.28
esmx-enus	37.7	50.6	52.90	69.40
engb-dede	38.54	52.38	53.88	61.91

Table 5: search metric (nDCG@16) of source queries and query MT and reference translations

Language pair	**lower bound**	**upper bound**	**impact range**
ptpt-eses	42.53	55.65	13.12
frca-enca	33.34	53.47	20.13
engb-dede	38.54	61.91	23.37
nlnl-dede	31.11	56.76	25.65
enus-jajp	36.20	62.30	26.10
esmx-enus	37.70	69.40	31.70
ptbr-enus	27.20	60.28	33.08
trtr-dede	26.80	63.90	37.10
kokr-enus	19.59	60.42	40.83
dede-enus	17.78	60.27	42.49
mlin-enin	4.00	58.28	54.28
knin-enin	2.90	58.28	55.38

Table 6: The MT impact range (nDCG@16)

		MT launching impact	
Rank	**impact range**	**ΔnDCG/ ΔBleu**	**ΔnDCG/ ΔChrf**
1	knin-enin	mlin-enin	dede-enus
2	mlin-enin	kokr-enus	knin-enin
3	dede-enus	trtr-dede	mlin-enin
4	kokr-enus	knin-enin	ptbr-enus
5	trtr-dede	dede-enus	kokr-enus
6	ptbr-enus	ptbr-enus	trtr-dede
7	esmx-enus	esmx-enus	nlnl-dede
8	enus-jajp	nlnl-dede	esmx-enus
9	nlnl-dede	frca-enca	engb-dede
10	engb-dede	engb-dede	frca-enca
11	frca-enca	enus-jajp	enus-jajp
12	ptpt-eses	ptpt-eses	ptpt-eses

Table 7: Language pair ranking based on the MT launching impact

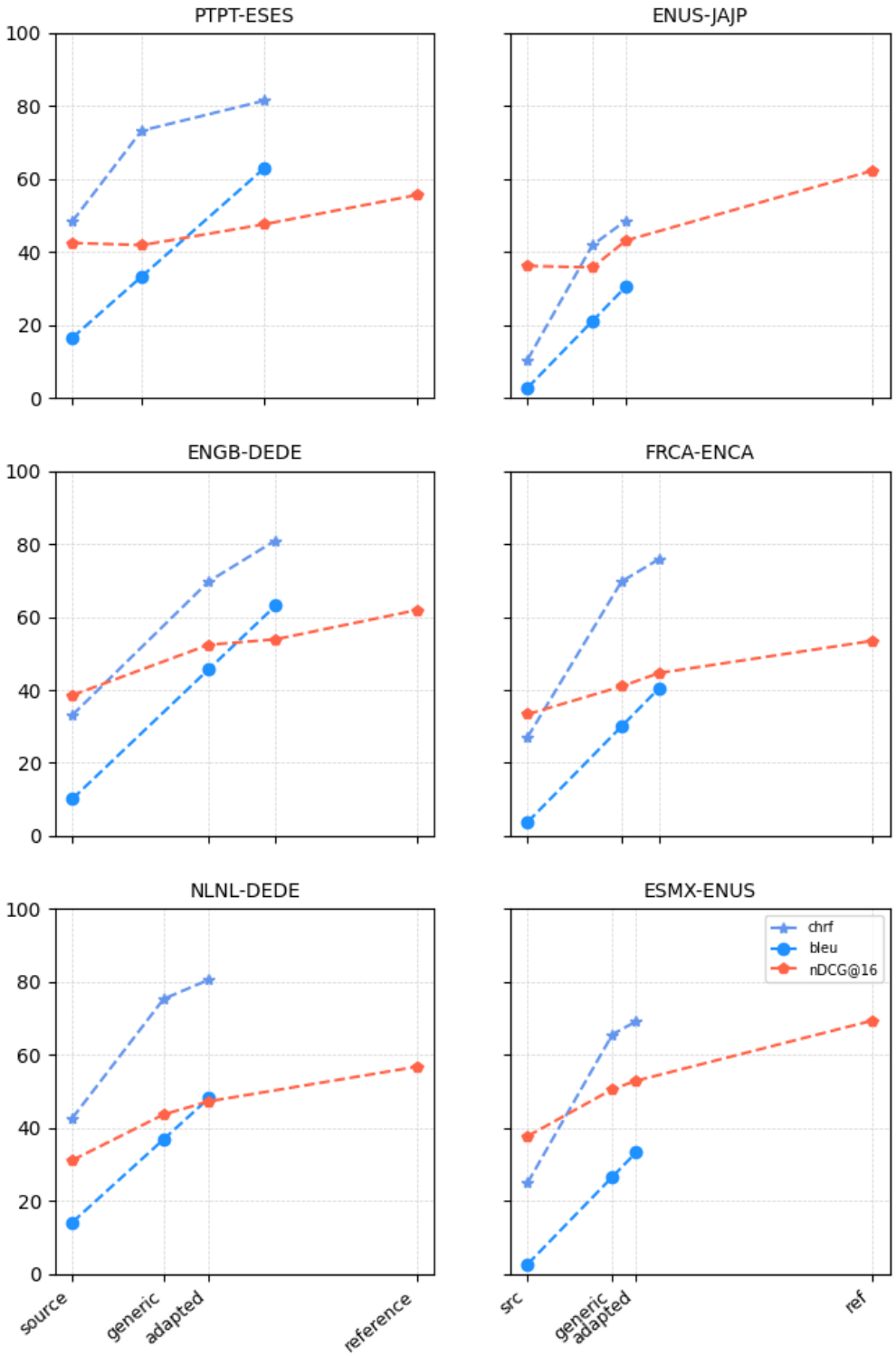

Figure 5: MT quality metrics and search metrics.png

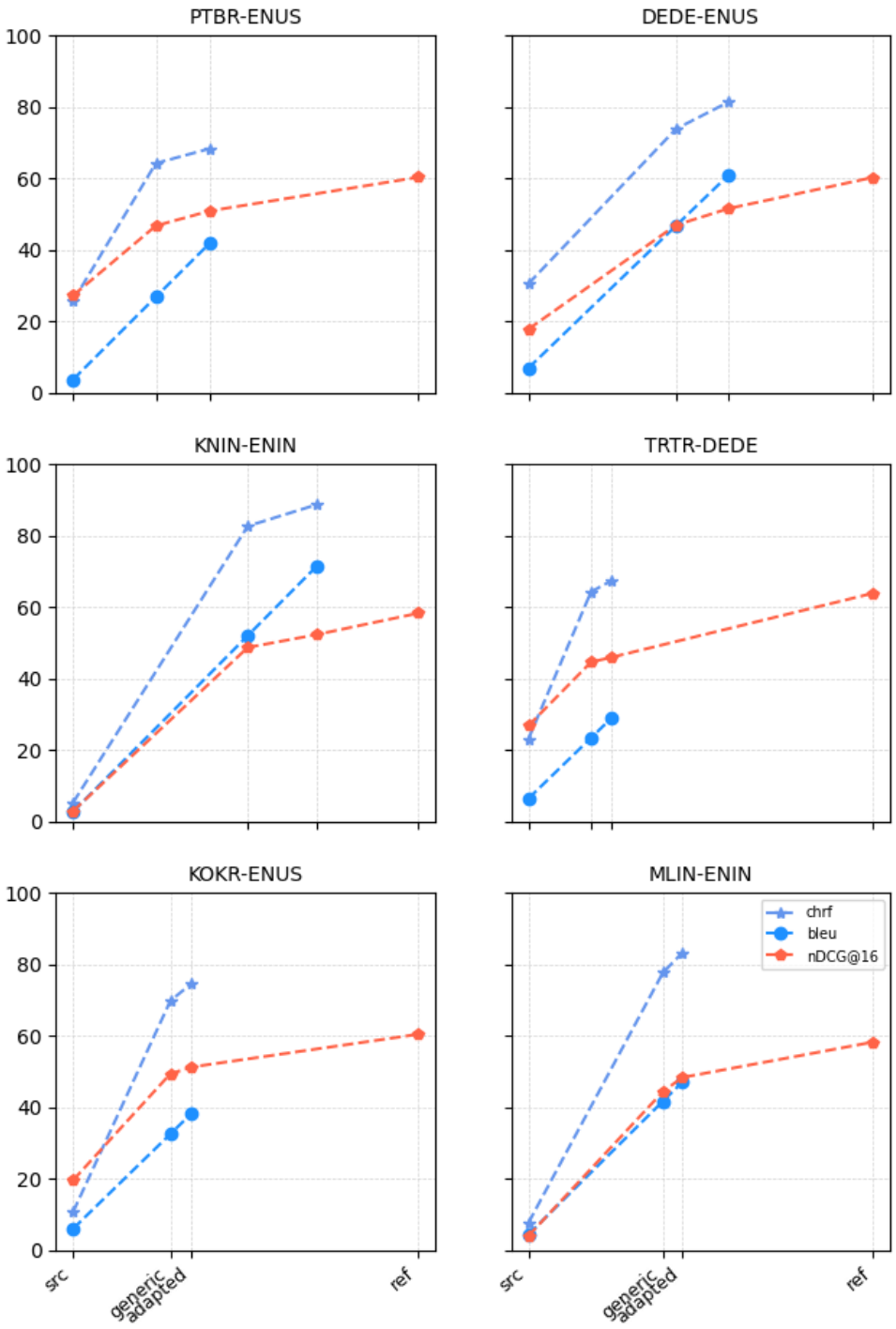

Figure 6: MT quality metrics and search metrics