Evaluating Machine Translation in Cross-lingual E-Commerce Search

Bryan Zhang bryzhang@amazon.com
Liling Tan lilingt@amazon.com
Amita Misra misrami@amazon.com
Amazon.com

Abstract

Multilingual query localization is integral to modern e-commerce. While machine translation is widely used to translate e-commerce queries, evaluation of query translation in the context of the down-stream search task is overlooked. This study proposes a search ranking-based evaluation framework with an edit-distance based search metric to evaluate machine translation impact on cross-lingual information retrieval for e-commerce search query translation, The framework demonstrate evaluation of machine translation for e-commerce search at scale and the proposed metric is strongly associated with traditional machine translation and traditional search relevance-based metrics.

1 Introduction

Multilingual search capability is essential for modern e-commerce product discovery (Lowndes and Vasudevan, 2021). Localization of e-commerce sites have led users to expect search engines to handle multilingual queries. Recent proposals of cross-lingual information retrieval handles multilingual queries and language-agnostic cross-borders product indexing have 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 machine translation (Nie, 2010; Rücklé et al., 2019; Saleh and Pecina, 2020; Bi et al., 2020; Jiang et al., 2020).

Machine translation (MT) is notoriously hard to evaluate manually; *human evaluation is slow, expensive and inconsistent* (Pierce and Carroll, 1966; Callison-Burch et al., 2007; Graham et al., 2013; Tan et al., 2015; Scarton and Specia, 2016; Freitag et al., 2021). Automated machine translation evaluation metrics have evolved from simple word error rates (Levenshtein et al., 1966; Tillmann et al., 1997) to modern string based metrics that usually ignores the source inputs and uses a single reference (Papineni et al., 2002; Doddington, 2002; Banerjee and Lavie, 2005; Popović, 2015). Neural evaluation metrics (Vela and Tan, 2015; Sellam et al., 2020; Thompson and Post, 2020; Rei et al., 2020) have gain recent popularity as they attempt to incorporate human annotations and multi-references through supervised learning. Although neural metrics have shown to agree more with human evaluation, they are built off language models that introduces new biases (Amrhein and Sennrich, 2022).

Despite the usefulness of evaluation metrics, machine translation is often used as interim application and objectives of the downstream tasks could have varying levels of tolerance of the inherent translation quality. Keeping the actual utility of machine translation in mind, extrinsic task-based evaluations were developed for spoken-language systems (Thomas, 1999; Akiba et al., 2004; Schneider et al., 2010; Anastasopoulos et al., 2021; Roy et al., 2021), information extraction (Sudo et al., 2004; Laoudi et al., 2006), automatic post-editing (Chatterjee et al., 2015, 2017) and domain-specific translation that requires different fidelity requirements (Cuong et al., 2016; Song et al., 2019; Li et al., 2020).

Information retrieval evaluation usually involves human-annotated relevance labels of search results candidates. Industrially, the scale of annotating a representative sample is impractical and can only serve as anecdotal evidence of search quality. As a proxy for human annotations, it is common to use behavioral signals from clicks and purchases (Wu et al., 2018). However, these behavioral signals pose a cold-start problem where such information is unavailable for newly established marketplaces.

In this paper, we examine the evaluation of machine translation of search query in the context of cross-lingual e-commerce search. We propose:

1. a **rank-based evaluation framework** to evaluate MT in Cross-lingual information retrieval (CLIR) through ranking-based search metrics using behavioral signals (from the marketplace of the target language) as a proxy to relevance information without any human annotation; this framework can be used to create for e-commerce CLIR test sets at scale.

2. a novel **edit-distance based metric** using Levenshtein edit distance to measure the divergence between the search results from machine translated queries and the search results from the human translated queries, this metric does not need any relevance information.

The rest of the paper is structured as follows. Section 2 gives an overview of the proposed ranking-based evaluation framework and edit-distance based evaluation framework for e-commerce Cross-lingual Information Retrieval (CLIR). Section 3 describes the experiment setup on the test set used to evaluate Machine Translation (MT) models tuned on search data and the edit-distance metric hyper-parameters. Section 4 presents our experiment results and analysis of the association between MT metric, traditional nDCG and the Levenshtein edit-distance based metric proposed in this paper. Section 5 presents related work and Section 6 concludes the paper.

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

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 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; and a novel edit-distance based framework to measure the difference in search candidate ranks between the human and machine translated queries without the need for relevance information.

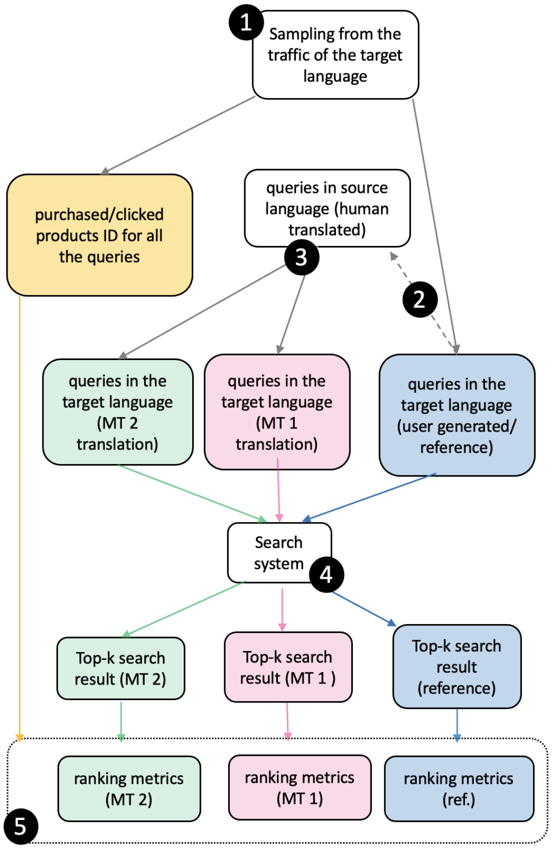
1
Sampling from the traffic of the target language
purchased/clicked products ID for all the queries
queries in source language (human translated)
3
2
queries in the target language (MT 2 translation)
queries in the target language (MT 1 translation)
queries in the target language (user generated/ reference)
Search system
4
Top-k search result (MT 2)
Top-k search result (MT 1)
Top-k search result (reference)
ranking metrics (MT 2)
ranking metrics (MT 1)
ranking metrics (ref.)
5

Figure 1: Ranking-based Evaluation Framework to Evaluate MT in E-commerce CLIR

To our best knowledge, there is no systematic study on cross-lingual information retrieval study for e-commerce search that neither requires ground-truth click/purchase information nor human annotated relevance data.

2.1 Ranking-based Evaluation Framework

Figure 1 illustrates the ranking-based evaluation framework to evaluate machine translation in the context of cross-lingual information retrieval for e-commerce.

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).[1] *We refer to these queries as* Q_{ref}.

   (a) To allow computation of traditional relevance metrics, record 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}.

[1] 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.

2. 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}.*

3. Translate the Q_{src} with the different MT models in consideration, e.g. MT1 and MT2 systems. *We refer to these machine translated queries as Q_{mt1} and Q_{mt2}.*

4. Search for the respective candidate products using the machine translated queries Q_{mt1}, Q_{mt2} and original Q_{ref}; retrieving top-k search results respectively, R_{mt1}, R_{mt2} and R_{ref}.

5. Use R_{mt1}, R_{mt2} and R_{ref} to directly compute edit-distance based evaluation metric (refer to section 2.2). If available, additionally use P_{id} and P_{freq} (as ground truth) with R_{mt1}, R_{mt2} and R_{ref} to compute traditional relevance based metrics such as nDCG.

We propose the above framework to evaluate machine translation in the context of Cross-lingual Information Retrieval (CLIR) for e-commerce queries. Using clicks and purchase relevance information P_{id} and P_{freq} as ground truth, we can compute an upper-bound for traditional search metrics from R_{ref}.

We provide an example of how an evaluation data can be created using the proposed ranking-based framework to evaluate a Spanish to English translation model:

Step 1: Given a sample query in the target language that the search index is built on, e.g. "*turn signal bulb*", Q_{ref}, we first extract the clicks and purchase product IDs associated with the query (P_{id} and P_{freq}).

Step 2: Next, we collect the human reference translation for the query "*foco para luz direccional*" and use that as (Q_{src})

Step 3: Then, we translate Q_{src} with MT1 and MT2 translation models, e.g. "*turn signal light bulb*" as Q_{mt1} and "*bulb for directional light*" as Q_{mt2}

Step 4: We put the translated queries, Q_{mt1} and Q_{mt2}, and the original English query, Q_{ref}, through the e-commerce search engine to retrieve the product search results R_{mt1}, R_{mt2} and R_{ref}

Step 5: Finally, we can compute the traditional search metrics, e.g. nDCG, with the R_{mt1}, R_{mt2} and R_{ref} using P_{id} and P_{freq} as relevance ground truth.

Using the search results from Step 4, the next section introduces the edit-distance based CLIR metric in additional to the traditional search metrics in Step 5.

2.2 Edit-distance based Evaluation metric without Relevance Information

In cold-start situations, clicks and purchases behavioral data is not available making it impossible to compute relevance based metric for machine translation in CLIR setting. Hence, we propose a novel edit-distance based evaluation framework using edit-distance based metric to approximate search performance without the need of relevance information.

Using the search results from the reference translation R_{ref} as a silver standard, we formulate the problem of measuring difference between the product candidates R_{ref} and R_{mt}. Edit-distance based similarity between R_{ref} and R_{mt} is computed by treating each product candidate like a character in a string. In short, we expect the search results of a good MT system not to diverge much from the search results produced by the reference translation.

We propose to use Levenshtein distance (Levenshtein et al. (1966)) to model this divergence. Levenshtein distance (Levenshtein et al., 1966) is widely used to measure string sequence difference, e.g. for string correction (Navarro, 2001). Any distance algorithm in effect

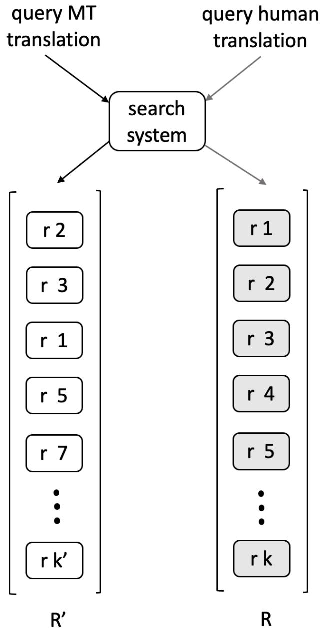

Figure 2: Edit-distance based Metric

would work but Levenshtein distance has more advantages over some of other distance metrics: we observe that it is common that some products are not shared by the two search results returned from different query translations of the same source query. Similarity metrics such as Spearman's rank correlation coefficient (Daniel (1990)) and Kendall Tau (Kendall (1938)) measure the ranking / ordering similarity of products only shared by two search results, Jaccard similarity coefficient (Jaccard (1912)) is a set-based similarity measure without consideration of the ranking/ ordering of the products in the search results. The Levenshtein distance (Levenshtein et al. (1966)) has the edit operations of insertions, deletions or substitutions, those edit operations can reflect both the products discrepancy between two search results, and also the ranking/ordering difference of the shared products in the search results.

Figure 2 illustrates the usage of the Levenshtein distance on search results. Formally, let R be the top-K search result returned from reference query translation and R' is from the MT query translation. Then we compute the edits (deletion, insertion and substitution) needed to make R' become R. Less edits indicates better search performance.[2]

3 Experimental Setup

Language pairs and locales: We select 4 language pairs from two e-commerce locales for our experiments, they are:

- Spanish-English (es-en) and Hebrew-English (he-en) in the US marketplace and
- English-German (en-de) and Polish-German (pl-de) in the German marketplace

Test data: The test data is created as described Step 1 and Step 2 from Section 2.1 as proposed in 2.1. The test set comprises 4000 queries per marketplace each translated into their respective

[2] The best possible score to achieve is when the $R' = R$ that results in a score of 0.0.

language pairs. To compare our proposed framework and metric with traditional relevance based evaluation, we also stored the purchased product IDs associated with the queries.

Machine Translation (MT) models: We trained two models per language pair to compare (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).

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 and Levenshtein edit-distance metric (Lev@16); the edit cost for Levenshtein is set uniformly at 1 for substitution, deletion, and insertion.

4 Results and Analysis

Language	**Model**	↑ **Bleu**	↑ **nDCG@16**	↓ **Lev@16**	**Upper Bound (nDCG@16)**
es-en	Generic	51.69	0.46	10.62	0.60
	Search	**54.04**	**0.53**	**10.23**	
he-en	Generic	48.25	0.43	11.49	0.60
	Search	**56.12**	**0.47**	**10.00**	
en-de	Generic	42.59	0.45	11.23	0.62
	Search	**63.08**	**0.54**	**7.91**	
pl-de	Generic	35.62	0.39	12.51	0.62
	Search	**56.24**	**0.48**	**9.49**	

Table 1: MT quality metrics, ranking metrics and distance-based metrics for all the MT models

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. Table 1 presents the traditional BLEU [3] machine translation evaluation metric, normalized discounted cumulative gain with top-16 search results (nDCG@16) with behavioral signal-purchased product IDs as a proxy to relevance for computation, and the proposed Levenshtein edit-distance based CLIR metric proposed in this paper (Lev@16).

As an upper-bound reference, Spanish to English (*es-en*) and Hebrew to English (*he-en*) achieve an nDCG@16 score of 0.60 when using the reference translation that produces the R_{ref} search results. Likewise, English to German (*en-de*) and Polish to German models (*pl-de*), they achieve an nDCG@16 score of 0.62 for their respective R_{ref}.

We can use these upper-bounds to expect the possible improvements that can be made to the machine translation in the cross-lingual IR setting. For example, *es-en* language pair has an 0.53 nDCG score while *he-en* in the same marketplace scores at 0.47, we can expect that there is more room for improvement for the *he-en*, given that the reference translation R_{ref} achieved a score of 0.60.

Juxtaposing the generic and search MT models, we expect the search models to perform better given the domain-specific tuning. The difference in machine translation performance as

[3] Sacrebleu version 2.0.0 (Post, 2018)

measured by BLEU in Table 1 is correspondingly reflected in the relevance-based and edit-distance based search metrics. Most notably, the Polish to German model differs in BLEU score for generic and search variants by over 20 BLEU and nDCG@16 improved by +0.09 and Lev@16 improved from 12.5 to 9.5, (25% improvements for both nDCG and Lev).

4.1 Correlation between MT, relevance-based and Edit-distance metric

In order to understand the proposed edit-distance based metric with regard to the MT and search metrics, we further conduct a correlation study of the following three metrics: BLEU, nDCG and Levenstein Distance.

The Pearson's R correlation values between the traditional machine translation metric (BLEU), relevance-based search metric (nDCG@16) and edit-distance based search metric (Lev@16) of the Search MT and Generic MT models are presented in Table 2 and 3.[4], the nDCG is scaled to 0-100 for the computation convenience. As Levenshtein measures of the divergence between the search results from human query translation and MT query translation, we use the absolute value of ΔnDCG between MT and human query translations for this correlation study.

	Search MT		
Language	BLEU / nDCG	BLEU / Lev	ΔnDCG / Lev
en-de	0.32	0.89	0.61
pl-de	0.36	0.88	0.60
es-en	0.38	0.88	0.58
he-en	0.41	0.89	0.59

Table 2: Pearson Correlation between MT and Search Metrics for Search MT Models

	Generic MT		
Language	BLEU / nDCG	BLEU / Lev	Δ nDCG / Lev
en-de	0.33	0.86	0.56
pl-de	0.38	0.84	0.53
es-en	0.39	0.88	0.57
he-en	0.41	0.86	0.56

Table 3: Pearson Correlation between MT and Search Metrics for Generic MT Model

We can interpret the above correlation values as the mean cross-product of the standardized MT and search metrics (Lee Rodgers and Nicewander, 1988), values closer to 1.0 reflects correlation between the metrics and values closer to 0.0 indicates disassociation. Similar to Sloto et al. (2018), we find that BLEU does not correlate with nDCG improvements. However, we find it interesting that BLEU is strongly correlated to Levenshtein metric that demonstrates that higher BLEU values would correspond to lower Levenshtein distance and vice versa. Moreover, ΔnDCG has a moderate positive correlation to Levenshtein distance. Therefore, Levenstein distance can be an effective approximate metric for the search performance of query translation when it is impossible to compute the rank-based search metrics such as nDCG.

[4] As Levenshtein metric is inversely related to BLEU, i.e. lower Lev is better and higher BLEU is better, we multiply Lev with coefficient -1 before computing Pearson R.

4.2 Edit-distance metric with Varying K

Search engines adjust the number of top-K search results for different applications. Within e-commerce search, there are also varying K values implemented for practical reasons. For example, sponsored search results have limited real estate on the site, thus sponsored search has small values of K; normal product search has more allowance for larger k values. We investigate how edit-distance based search metrics differs with varying K search results.

k	es-en	he-en	en-de	pl-de
4	2.39	2.33	1.82	2.21
8	4.95	4.84	3.82	4.59
16	10.23	10.00	11.49	9.49
100	66.52	65.3	50.66	61.31

Table 4: Levenshtein Metric of Search MT models with different top-K search results

As the number of search candidates increases, we expect the distance between the R_{mt} to diverge from the R_{ref} and metric scores would linearly to the K value. Table 4 presents the Levenshtein metric results for the Search MT models with varying top-K search results.

	Generic MT			
Language	ΔnDCG@4 /Lev@4	ΔnDCG@8 /Lev@8	ΔnDCG@16 /Lev@16	ΔnDCG@100 /Lev@100
pl-de	0.56	0.55	0.53	0.49
en-de	0.59	0.58	0.56	0.52
he-en	0.60	0.59	0.56	0.52
es-en	0.61	0.59	0.57	0.53

Table 5: Pearson Correlation between Levenstein distance and ΔnDCG for Generic MT

	Search MT			
Language	ΔnDCG@4 /Lev@4	ΔnDCG@8 /Lev@8	ΔnDCG@16 /Lev@16	ΔnDCG@100 /Lev@100
pl-de	0.63	0.62	0.60	0.57
en-de	0.63	0.62	0.61	0.57
he-en	0.62	0.61	0.59	0.55
es-en	0.61	0.60	0.58	0.54

Table 6: Pearson Correlation between Levenstein distance and ΔnDCG for Search MT

As top-4, top-8, top-16 are commonly used for top-K search result evaluation for the cross-lingual E-commerce search, we also conduct a correlation study for the ΔnDCG and Levenstein distance with varying K as Table 5 and 6. The experiment setup is identical to the correlation study in section 4.1. As K increases, the correlation slightly decreases for both search and generic MT. We observe that there is subtle distinction in correlation in the range of $K \leq 16$; For search MT, there is moderate positive correlation between the ΔnDCG and Levenstein distance when $K \leq 16$. It further shows that Levenstein distance can be an effective approximate metric to evaluate the search performance of query translation in various cross-lingual E-commerce search scenarios when it is impossible to compute the rank-based search metrics.

5 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 esp. 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).

6 Conclusion

In this study, we introduce a framework which provides a recipe to evaluate machine translation in the context of cross-lingual e-commerce search at scale. Additionally, we proposed an edit-distance based metric `Lev@K` to evaluate MT quality that bypasses the reliance on behavioral signals and/or expensive and slow relevance annotations from human.

The proposed metric has shown correlations with traditional relevance-based search metric and it is also strongly associated with the classic machine translation evaluation metric. The difference between a machine translation system as measured by BLEU can be demonstrated with the proposed edit-distance based metric in the context of cross-lingual search. We suggest the use of the `Lev@K` metric in future CLIR researches in addition to the traditional search metrics, especially when relevance information is unavailable.

References

Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., and Tsujii, J. (2004). Overview of the iwslt04 evaluation campaign. In *IWSLT*, pages 1–12.

Amrhein, C. and Sennrich, R. (2022). Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet.

Anastasopoulos, A., Bojar, O., Bremerman, J., Cattoni, R., Elbayad, M., Federico, M., Ma, X., Nakamura, S., Negri, M., Niehues, J., Pino, J., Salesky, E., Stüker, S., Sudoh, K., Turchi, M., Waibel, A., Wang, C., and Wiesner, M. (2021). FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

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

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chatterjee, R., Gebremelak, G., Negri, M., and Turchi, M. (2017). Online automatic post-editing for MT in a multi-domain translation environment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 525–535, Valencia, Spain. Association for Computational Linguistics.

Chatterjee, R., Weller, M., Negri, M., and Turchi, M. (2015). Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China. Association for Computational Linguistics.

Cuong, H., Frank, S., and Sima'an, K. (2016). ILLC-UvA adaptation system (scorpio) at WMT'16 IT-DOMAIN task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 423–427, Berlin, Germany. Association for Computational Linguistics.

Daniel, W. (1990). *Applied Nonparametric Statistics*. Duxbury advanced series in statistics and decision sciences. PWS-KENT Pub.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Domhan, T., Denkowski, M., Vilar, D., Niu, X., Hieber, F., and Heafield, K. (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.

Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (2021). Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Hui, K., Yates, A., Berberich, K., and de Melo, G. (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.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.

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

Jiang, Z., El-Jaroudi, A., Hartmann, W., Karakos, D., and Zhao, L. (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.

Kendall, M. G. (1938). A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Laoudi, J., Tate, C. R., and Voss, C. R. (2006). Task-based MT evaluation: From who/when/where extraction to event understanding. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Li, J., Liu, C., Bing, L., Liu, X., Li, H., Wang, J., Zhao, D., and Yan, R. (2020). Cross-lingual low-resource set-to-description retrieval for global e-commerce. *ArXiv*, abs/2005.08188.

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

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

Lu, H., Hu, Y., Zhao, T., Wu, T., Song, Y., and Yin, B. (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.

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

McDonald, R., Brokos, G., and Androutsopoulos, I. (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.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

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

Nigam, P., Song, Y., Mohan, V., Lakshman, V., Ding, W. A., Shingavi, A., Teo, C. H., Gu, H., and Yin, B. (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.

Nigam, P., Song, Y., Mohan, V., Lakshman, V., Weitian, Ding, Shingavi, A., Teo, C. H., Gu, H., and Yin, B. (2019b). Semantic product search.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pierce, J. R. and Carroll, J. B. (1966). Language and machines: Computers in translation and linguistics. In *A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.*, page 124. National Research Council.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Post, M. (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.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Roy, S., Brunk, C., Kim, K.-Y., Zhao, J., Freitag, M., Kale, M., Bansal, G., Mudgal, S., and Varano, C. (2021). Using machine translation to localize task oriented nlg output. *arXiv preprint arXiv:2107.04512*.

Rücklé, A., Swarnkar, K., and Gurevych, I. (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.

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

Saleh, S. and Pecina, P. (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.

Sas, C., Beloucif, M., and Søgaard, A. (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.

Scarton, C. and Specia, L. (2016). A reading comprehension corpus for machine translation evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3652–3658, Portorož, Slovenia. European Language Resources Association (ELRA).

Schneider, A. H., van der Sluis, I., and Luz, S. (2010). Comparing intrinsic and extrinsic evaluation of MT output in a dialogue system. In *Proceedings of the 7th International Workshop on Spoken Language Translation: Papers*, pages 329–336, Paris, France.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sloto, S., Clifton, A., Hanneman, G., Porter, P., Gates, D., Hildebrand, A. S., and Kumar, A. (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.

Song, K., Zhang, Y., Yu, H., Luo, W., Wang, K., and Zhang, M. (2019). Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Sudo, K., Sekine, S., and Grishman, R. (2004). Cross-lingual information extraction system evaluation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 882–888, Geneva, Switzerland. COLING.

Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., 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.

Tan, L., Dehdari, J., and van Genabith, J. (2015). An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Thomas, K. (1999). Designing a task-based evaluation methodology for a spoken machine translation system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 569–572, College Park, Maryland, USA. Association for Computational Linguistics.

Thompson, B. and Post, M. (2020). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated dp based search for statistical translation. In *In European Conf. on Speech Communication and Technology*, pages 2667–2670.

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

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

Vela, M. and Tan, L. (2015). Predicting machine translation adequacy with document embeddings. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 402–410, Lisbon, Portugal. Association for Computational Linguistics.

Wu, L., Hu, D., Hong, L., and Liu, H. (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.

Zhang, L., Rettinger, A., Färber, M., and Tadić, M. (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.