# Enhancing Arabic Machine Translation for E-commerce Product Information: Data Quality Challenges and Innovative Selection Approaches

**Bryan Zhang**
Amazon
bryzhang@amazon.com

**Salah Danial**
Amazon
sdanials@amazon.ae

**Stephan Walter**
Amazon
sstwa@amazon.de

## Abstract

Product information in e-commerce is usually localized using machine translation (MT) systems. The Arabic language has rich morphology and dialectal variations, so Arabic MT in e-commerce training requires a larger volume of data from diverse data sources; Given the dynamic nature of e-commerce, such data needs to be acquired periodically to update the MT. Consequently, validating the quality of training data periodically within an industrial setting presents a notable challenge. Meanwhile, the performance of MT systems is significantly impacted by the quality and appropriateness of the training data. Hence, this study first examines the Arabic MT in e-commerce and investigates the data quality challenges for English-Arabic MT in e-commerce then proposes heuristics-based and topic-based data selection approaches to improve MT for product information. Both online and offline experiment results have shown our proposed approaches are effective, leading to improved shopping experiences for customers.

## 1 Introduction

As e-commerce shopping websites are localized worldwide, customers now are provided with options to browse products in their preferred languages other than the primary language of the store. For instance, customers from the Kingdom of Saudi Arabia (KSA) can shop in both English and Arabic in the KSA store. Modern e-commerce stores provide multi-lingual product discovery (Rücklé et al., 2019; Nie, 2010; Saleh and Pecina, 2020; Bi et al., 2020; Jiang et al., 2020; Lowndes and Vasudevan, 2021), and product information such as titles, descriptions, and bulletpoints are usually translated using machine translation (MT) systems (Way, 2013; Guha and Heger, 2014; Zhou et al., 2018; Wang et al., 2021). Product information in e-commerce demands accurate, culturally relevant, and contextually appropriate translations, which has significant impact on the customers' shopping experiences. The highly complex morphology of Arabic as well as other linguistic aspects have made the machine translation from and to Arabic a lot more challenging (Ameur et al., 2020; Alkhatib and Shaalan, 2018). Moreover, the multitude of dialectal variants along social and geographic dimensions introduce further linguistic challenges to MT (Habash, 2010). Hence in order to train Arabic MT systems in the e-commerce industrial setting, typically a larger volume of training data needs to be acquired from a wider range of data sources to address the complexity of the Arabic language. Moreover, as e-commerce product catalogs continue to expand, the task of maintaining up-to-date machine translation systems poses significant challenges. When the vast amount of product information is sourced from various sellers or suppliers, each can present the data differently. As a result, the inconsistencies and noise in the source data can have a negative impact on MT systems. Meanwhile, validating a substantial volume of data for MT training at scale becomes increasingly difficult and time-consuming, demanding significant resources for manual review and error correction to guarantee the accurate interpretation of product information.

Therefore, in this study, we first investigate the **training data quality issues and challenges** for Arabic MT in e-commerce, and identify two major data issue patterns based on our observations and addressing the data quality challenges from the periodic data acquisition. Then we propose **heuristics-based** and **topic-based data selection approaches** for Arabic MT. The heuristics-based data selection approach leverages the identified data issue patterns that are typical to the Arabic training data in e-commerce and proposes straightforward and effective data filters to remove the undesirable noisy data for training data quality



*Proceedings of the The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, pages 150–157

improvement; The topic-based data selection approach first clusters the data based on the textual patterns then choose the clusters of the clean data for MT training so that the data of new and unknown noise patterns from the periodic data sourcing can be removed. We experiment our proposed approaches separately and in combination for the case study of English-Arabic MT. The offline experiment results have shown that the application of two approaches in combination can further improve the MT by 4.47% for BLEU on average across three domains (product titles, descriptions and bulletpoints), and 9.32% for BLEU for titles. The online A/B experiment results further have shown the customers' shopping experiences have been improved, which indicates the effectiveness of our proposed approaches.

## 2 Training data for Arabic MT in e-commerce

### 2.1 Arabic language in e-commerce

Arabic language is rich in morphology and has a large number of dialects given an Arabic-speaking region, hence *Modern Standard Arabic* (MSA) is usually a practical choice for the Arabic MT in e-commerce. Unlike regional dialects, MSA Arabic is understood by the majority across the Arab world, providing a unified platform for communication. In the context of e-commerce, this is particularly advantageous as it enables us to effectively convey our product titles, descriptions and bulletpoints in a consistent manner. On the other hand, we have also observed that it is beneficial to adapt MSA to some extent for specific regions. For instance, the word *case* in the *iPhone 14 pro max transparent case with stand Dual 360° Rotating ring* has a more formal MSA translation غطاء. However, when the translation is used specifically for the store in Egypt, the dialectal variation جراب for the word *case* is preferred since we observe it can improve customers' shopping experience.

### 2.2 Common Arabic data issues in e-commerce

**Many-to-one and one-to-many cases**: we have observed that it is more common in the Arabic data that some source texts have multiple target texts (reference translations), particularly for language pairs where the target language is Arabic.[1] Those multiple target variants can be either translation or transliteration. For example, given the source *Stainless Steel*, there are target texts فولاذ مقاوم للصدأ (translation) and ستانلس ستيل (transliteration); they can also be the dialectal variations in Arabic, For example: given source text *Cases and Covers*, the target texts can be جرابات وحافظات حماية or كوفرات وجرابات; It is also possible that the multiple targets are just inaccurate translations, for example: *Product colour: Silver* can have more than one inaccurate target translation such as لون المنتج: فضة. الوزن: ٤٨٩ غرام and لون المنتج: فضة. الوزن: ٥٨ غرام.

**Incorrect languages**: Given the wide range of the data sources for data acquisition, it is common to have noisy data acquired in a language that is not part of the language pair. We have observed that for Arabic data, such noisy texts can be entirely in a different language or also often in mixed languages such as partial English and Arabic, which poses challenges for existing language detection tools that are tailored for texts usually in one language.

### 2.3 Emerging new noise patterns

Product catalogs continue to expand in the dynamic e-commerce, therefore, it is crucial to acquire newer data periodically to update the MT systems. Considering the rich morphology and dialectal variations of Arabic, the vast amount of product information is often acquired from a larger number of sellers or suppliers, and each of which can present the data differently. As a result, inconsistencies and noise emerge inevitably during each data acquisition cycle in the source data, which can have a negative impact on MT systems. Although we are aware of the various common noise patterns and data issues, it is challenging to detect such new noise patterns or data issues given the quality of the data and the complexity of the Arabic language.

## 3 Heuristics-based data selection approach

**1:M/N:1 data filter**: When the source (or target) texts have a larger number of target (or source)

[1] Some target texts have multiple source texts, particularly for language pairs where the source language is Arabic.

texts, it is challenging to validate the quality of such data at scale. When a larger number of variants can be mapped to a single source or target texts, it is also more likely that such data can be defected data and have a negative impact on the MT training. Therefore, we propose a heuristics-based **1:M/N:1 data filter**. $M$ refers to the number of target references for a given source text whereas $N$ refers to the number of the source texts given a target in the training data. We can use this filtering mechanism to detect and remove the data which have a larger number of mapped source or target texts than $M$ and $N$ respectively.

**Script-based language filter**: We propose a straightforward **Script-based language filter** for language pairs involving Arabic to filter the data that is not in the expected language. This script-based language filter is based on the string overlapping between an input string and the alphabet set of the given language. As Arabic language is morphologically different from most languages, such filtering mechanism can be effective. We apply this filtering mechanism to detect the language based on the ratio of the number of characters in a given string that belongs to the alphabet of the given language and the total number of characters in the input string. Given an input string $S$, $L$ is list of the letters/characters of input string $S$ ($|S| = |L|$), $A$ is the alphabet set of the given language, we define the filter ratio $T$ as equation 1

$$T = \frac{|S_{alphabet}|}{|S|} \tag{1}$$

where, $S_{alphabet} =< l_1, l_2 ... l_n >$ is a list of the letters $l_i (l_i \in S_{alphabet})$ where $l_i \in S$ and $l_i \in A$. This filtering mechanism can achieve a high precision especially when we decrease our filter ratio threshold ($T$) to make sure we only remove sentences with a large number of characters that do not belong to the expected character set.

## 4 Topic-based data selection

### 4.1 Topical clustering

We use Dirichlet Multinomial Mixture (DMM) (Nigam et al., 2000) and Collapsed Gibbs Sampling (CoGS) (Yin and Wang, 2014) for topical clustering. DMM and CoGS are efficient clustering algorithms capitalizing on symbolic text representation, making them ideal to cluster industry scale e-commerce data based on textual patterns. Moreover, the number of topic clusters is automatically inferred to adequately capture both frequent and rare textual patterns.

We use the DMM model to label each document (input text) with one topic tag. DMM is a probabilistic generative model for documents and embodies two assumptions about the generative process: first, the documents are generated by a mixture model; second, there is a one-to-one correspondence between mixture components and clusters. When generating document $d$, DMM first selects a mixture component (topic cluster) $k$ according to the mixture weights (weights of clusters) $P(z = k)$. Then document $d$ is generated by the selected mixture component (cluster) from distribution $P(d|z = k)$. We can characterize the likelihood of document $d$ with the sum of the total probability over all mixture components:

$$P(d) = \sum_{k=1}^{K} P(d|z = k)P(z = k) \tag{2}$$

where, $K$ is the number of mixture components (topic clusters). DMM assumes that each mixture component (topic cluster) is a multinomial distribution over words and each mixture component (topic cluster) has a Dirichlet distribution prior:

$$P(w|z = k) = P(w|z = k, \Phi) = \phi_{k,w} \tag{3}$$

$$P(z = k) = P(z = k|\Theta) = \theta_k \tag{4}$$

where, $\sum_{w}^{V} \phi_{w,k} = 1$ and $P(\Phi|\vec{\beta}) = Dir(\vec{\theta}|\vec{\beta})$ and $\sum_{k}^{K} \theta_k = 1$ and $P(\Theta|\vec{\alpha}) = Dir(\vec{\theta}|\vec{\alpha})$.[2]

The collapsed Gibbs sampling is used to estimate DMM parameters, documents are randomly assigned to $K$ clusters initially and the following information is recorded:

$\vec{z}$ is the cluster labels of each document

$m_z$ is the number of documents in each cluster $z$

$n_z^w$ is the number of occurrences of word $w$ in each cluster $z$

$N_d$ is the number of words in document $d$

$N_d^w$ is the number of occurrence of word $w$ in the document $d$

$V$ is the vocabulary of the corpus

[2]The weight of each mixture component (cluster) is sampled from a multinomial distribution which has a Dirichlet prior

The documents are traversed for a number of iterations. In each iteration, each document is reassigned to a cluster according to the conditional distribution of $P(Z_d = z|\vec{z}_{\neg d}, \vec{d})$, $\neg d$ means $d$ is not contained:

$$P(Z_d = z|\vec{z}_{\neg d}, \vec{d}) \propto \frac{m_{z,\neg d}+\alpha}{D-1+K\alpha} \frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(n^w_{z,\neg d}+\beta+j-1)}{\prod_{i=1}^{N_d}(n_{z,\neg d}+V\beta+i-1)} \quad (5)$$

where, hyper-parameter $\alpha$ controls the popularity of the clusters, hyper-parameter $\beta$ emphasizes on the similar words between a document and clusters.

### 4.2 Topic-based data selection

As Figure 1 shown, the data selection approach first clusters large volume of the training data. Empirically, larger clusters can capture the major topical and textual patterns so they are usually the clean desirable data whereas the smaller clusters can capture smaller and rare textual patterns so they are likely to be the noisy undesirable data. Additionally, we can also distinguish between desirable and undesirable data based on the data inspection of the clusters. Finally, only clusters of desirable data are chosen for training to improve MT. Data providers are also informed of the undesirable data patterns for future data quality control.

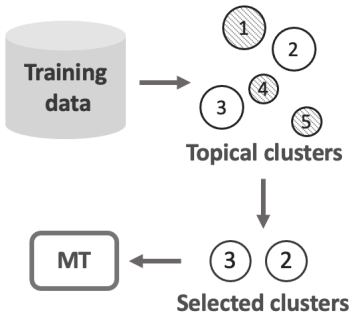


Figure 1: Choosing desirable data for MT training

## 5 Case study: English-Arabic MT

### 5.1 Experiment setup

**Data**: We train the MT models on a large volume of in-house generic training data and ∼20 million product-information data (product titles, descriptions and bulletpoints) for domain adaptation. We have three test data sets for product titles, descriptions and bulletpoints respectively. Each test data set has 2000 test segments and we evaluate the models using BLEU[3] and chrF (Popović, 2015) to assess the translation quality.

**Model**: We use the transformer-based architecture (Vaswani et al., 2017) with 20 encoder and 2 decoder layers with the Sockeye MT toolkit (Domhan et al., 2020) to train a generic MT using generic data and domain-specific data, then fine-tune the model on the domain-specific product information data for domain adaptation.

**Baseline Model**: The baseline MT model is first trained using generic data and domain-specific data, then is fine-tuned on the domain-specific product information data.

**Topic Clustering**: For the topic clusters, the source text is lower-cased, tokenized and stemmed using NLTK ToolKit (Bird et al., 2009), stemmed tokens with document frequency less than or equals to 2 are removed in the preprocessing steps. The initial upper-bound number of topical clusters is set to 500. The number of the topic clusters is inferred automatically during the collapsed Gibbs sampling process. The number of iterations is set to 30, and both hyper-parameters $\alpha$ and $\beta$ are set to 0.1.

We create 2-D plots using Jensen-Shannon distance (Fuglede and Topsoe, 2004) and multi-dimensional scaling technique (Borg and Groenen, 2005) with *LDAvis* (Sievert and Shirley, 2014) to easily visualize the size and relations of the topic clusters returned from the algorithm, and to inspect the topic words extracted from the clusters.

**Data filters**: 1:M/N:1 data filter: We choose $m$=$n$=10 and $m$=$n$=5 for the 1:M/N:1 data filter respectively. The former is more relaxed since each sentence can have up to 10 variants whereas the latter with $m$=$n$=5 is more strict.

**Language detection filter**: For the script-based language filter, we choose $T$=0.1, so the data will be removed if 10% or less of the sentence characters belong to the character set. We apply this language filter on both source and target texts. The character set for the source side was Latin (ISO-8859-1) and for the target side was Arabic

[3]SacreBLEU version 2.0.0 (Post, 2018)

(ISO-8859-6). We also incorporate two existing language detectors *Cybozu* language detection library[4] (Nakatani, 2010) and *FastText* (Joulin et al., 2016b,a) in addition to our script-based language filter.

### 5.2 Experiment results and analysis

#### Clustering Results

| **Indomain data size -(TTL/BP/DESC)** | ∼20 million |
|---|---|
| Num of total clusters | 374 |
| Num of major clusters (>1000 seg.) | 110 |
| Num of minor clusters | 264 |
| minor clusters % total data | 1.32% |

Table 1: Clustering result for the in-domain data (English data) for the bilingual indomain data for EnUs-ArAe

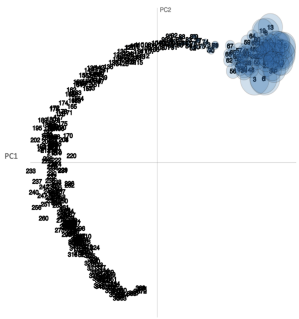


Figure 2: Plot of all the topic clusters with Principal Coordinate Analysis (PCoA)

Table 1 shows the clustering results using the source text of the ∼20 million indomain product data which includes titles (TTL), bulletpoints (BP) and descriptions (DESC). In total, there are 374 clusters extracted. We empirically consider clusters having 1000 segments or more data points as major clusters while those having less than 1000 segments as minor clusters. The total data from the minor clusters account for 1.32% of the total indomain training data.

We also generate 2-D data visualization as Figure 2 with projected clusters using Jensen-Shannon distance (Fuglede and Topsoe, 2004) and multi-dimensional scaling techniques such as Principal Coordinate Analysis (PCoA) (Borg and Groenen, 2005). We can use the plot to understand the relations of clusters, the sizes of the clusters are proportional to the size of the data assigned to the cluster. The long tail in the plot are all the minor clusters which deviate from the major clusters.

[4]https://github.com/shuyo/language-detection

#### Training Data Filtering results

| **Domain** | | **m=10, n=10** | **m=5, n=5** |
|---|---|---|---|
| TTL | **BLEU** | +0.73% | **+1.68%** |
| | **chrF** | -0.09% | **+1.98%** |
| DESC | **BLEU** | **+0.58%** | +0.27% |
| | **chrF** | -0.24% | **-0.16%** |
| BP | **BLEU** | **+0.56%** | +0.49% |
| | **chrF** | +0.00% | **+0.15%** |

Table 2: Quality improvement % of the model trained with 1:M/N:1 filter in the data selection over the baseline model trained with data without the filter (Configurations: $m$=$n$=5 and $m$=$n$=10

1:M/N:1 Filter: Previously, we have conducted a separate experiment with different $m$ and $n$ configurations using an older version of indomain data. We use two configurations $m$=$n$=5 and $m$=$n$=10 to filter the indomain data for MT training. As Table 2 shows, we have seen the average BLEU score and ChrF are improved by 0.64% and 0.51% respectively across three domains with the configuration of $m$=$n$=10. meanwhile, using a strict filter configuration of $m$=$n$=5 yields higher MT quality scores.

| **Domain** | | **Script Filter** | **Script Filter + Cybozu** | **Script Filter + FastText** |
|---|---|---|---|---|
| TTL | **BLEU** | **+2.94%** | -0.52% | +0.16% |
| | **chrF** | +1.61% | **+2.15%** | +0.43% |
| DESC | **BLEU** | -2.64% | -3.49% | **-2.56%** |
| | **chrF** | +0.38% | +0.38% | +0.38% |
| BP | **BLEU** | **+0.85%** | -1.47% | -0.23% |
| | **chrF** | **+0.91%** | +0.45% | +0.76% |

Table 3: Quality improvement % of the model trained with different language detection filters in the data selection over a baseline model without the filter.

**Language detection filter**: Table 3 shows the BLEU and chrF improvements over 3 domains of test set (TTL, DESC and BP) compared to a baseline trained using the latest indomain product information data. Using the script-based filter alone can improve the MT by 0.38% and 0.97% for the average BLEU score and ChrF, respectively. The experiment results have also shown that existing language detectors do not show substantial advantages to the data filtering on in addition to the straightforward script-based language detector.

| Domain | | HEU | TOPIC | HEU +TOPIC |
|---|---|---|---|---|
| TTL | **BLEU** | +0.93% | +7.20% | **+9.32%** |
| | **chrF** | +0.47% | +2.79% | **+3.83%** |
| DESC | **BLEU** | +1.31% | **+1.45%** | -0.02% |
| | **chrF** | **+0.95%** | +0.87% | **+0.95%** |
| BP | **BLEU** | **+4.10%** | +0.57% | **+4.10%** |
| | **chrF** | +2.05% | +0.41% | **+2.26%** |

Table 4: Quality improvement % of the model trained with both Heuristics-based (HEU) and Topic-based (TOPIC) data selection approaches compared with the baseline model trained with latest indomain data.

Furthermore, we have also conducted the experiment with both the heuristics-based (HEU) and topic-based (TOPIC) data selection approaches in combination. For the heuristics-based approach, we use the 1:M/N:1 data filter with configuration of $m$=$n$=5 as it yields better results in a separate study as discussed in Table 2, and we use our proposed script-based filter to remove data that is not English or Arabic. For the topic-based approach, we use the data from the major clusters as discussed in Table 1 for the MT model training. Then we apply the heuristics-based data selection approach to the data from the major clusters and use the filtered data to train an MT model.

Table 4 shows the MT quality metrics with both approaches alone and in combination using the aforementioned experimental configuration. We can see the MT model (HEU+TOPIC) with both approaches is further improved by 4.47% and 2.35% for BLEU and chrF on average across three domains (product titles, descriptions and bullet-points), and it also shows large improvement for titles by 9.32% and 3.83% for BLEU and chrF.

### 5.3 Human Evaluation and AB Testing

We have also conducted human evaluation for the MT translation quality in addition to the automatic metrics reported in the previous section, we provide human raters with hundreds of translations from the baseline MT and the newer MT (HEU +TOPIC) trained with both proposed approaches in combination, and let human raters assess the fluency and the adequacy of the translations, the newer MT's fluency and adequacy are improved by 3.1% and 3.29% compared with the baseline model.

As the Table 5 shows, in the example 1 the baseline model translated *sweet* to the sweets as candies whereas the newer model translates it better through transliteration since in Arabic such terms are not existent. In the example 2, baseline model incorrectly translates *brand new* to *new brand* whereas the newer model translates to *completely new* correctly.

| **Example 1** | |
|---|---|
| Source | *Great for Party Favors, Sweet 15 or 16* |
| Baseline | رائعة لهدايا الحفلات، حلوة ١٥ أو ١٦ |
| Newer | رائعة لهدايا الحفلات، سويت ١٥ أو ١٦ |
| **Example 2** | |
| Source | *Brand New And High Quality* |
| Baseline | العلامة التجارية الجديدة وعالية الجودة |
| Newer | جديد تمامًا وعالي الجودة |

Table 5: Translation examples from the baseline MT and newer MT (HEU+TOPIC)

We have further conducted online A/B testing in the Kingdom of Saudi Arabia (KSA) store with the English-Arabic MT. For the A/B testing, customers shopping in Arabic are presented with two different versions of the product information translations (titles, descriptions and bullet points) from the baseline model and the newer MT model (HEU +TOPIC) trained with heuristics-based and topic-based data selection approaches in combination. After a 4-week A/B testing experiment, the results have shown that the translations from the newer MT trained with our proposed approaches have a much larger positive impact on the customers' shopping experiences. This indicates the effectiveness of our approach.

## 6 Related Work

There are studies related to data selection for machine translation systems. (Mohiuddin et al., 2022) focuses on data selection for curriculum training through fine-tuning MT model on a selected by both deterministic scoring, (van der Wees et al., 2017) proposes dynamic data selection which varies the selected subset of training data between different training epochs to improve neural MT. Previous studies also have successfully used topic models to improve statistical machine translation (Eidelman et al., 2012; Hu et al., 2013; Xiong et al., 2015; Mathur et al., 2015) and neural machine translation (Zhang et al., 2016; Chen et al., 2019). (Mathur et al., 2015) integrates topic

models as feature functions in the phrase-tables to improve statistical machine translation for e-commerce domain adaption. (Zhang et al., 2016) presents an approach using topic models to increase the likelihood of word selection from the same topic as the source context. Instead of explicitly affecting the parameters or vocabulary selection, in this paper, we utilize a topical cluster model for data selection.

## 7 Conclusion

In this study, we first review and investigate the data quality validation challenges the Arabic machine translation systems for product information translation in e-commerce, Arabic language has rich morphology and dialectal variations, which can cause more data quality issues that are unique to acquired training data for developing MT translating from and to Arabic. Then we propose heuristics-based and topic-based data selection approaches to select clean and desirable data for neural MT training. Both offline experiment results and human evaluation have shown both approaches can improve the English-Arabic MT for product information. On-line A/B testing also shows customers' shopping experience has been improved with the translation from the MT trained with two approaches, which it shows the effectiveness of our proposed approaches.

## Limitations

In this study, we have proposed the approaches and conducted experiments for developing and improving English-Arabic MT for product information translation in e-commerce, and analyzed the offline MT translation quality and business impact. However, this study only focuses on the domain of e-commerce and the business case study of English-Arabic MT. In future work, we are planning to apply our proposed approaches to more language pairs involving Arabic and experiment with domains beyond product information.

## References


Manar Alkhatib and Khaled Shaalan. 2018. The key challenges for arabic machine translation. *Intelligent Natural Language Processing: Trends and Applications*, pages 139–156.

Mohamed Seghir Hadj Ameur, Farid Meziane, and Ahmed Guessoum. 2020. Arabic machine translation: A survey of the latest trends and challenges. *Computer Science Review*, 38:100305.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Ingwer Borg and Patrick JF Groenen. 2005. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2019. Neural machine translation with sentence-level topic context. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1970–1984.

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.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 115–119.

B. Fuglede and F. Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, pages 31–.

Jyoti Guha and Carmen Heger. 2014. Machine translation for global e-commerce on ebay. In *Proceedings of the AMTA*, volume 2, pages 31–37.

Nizar Habash. 2010. Introduction to arabic natural language processing. In *Introduction to Arabic Natural Language Processing*.

Yuening Hu, Ke Zhai, Vladimir Edelman, and Jordan Boyd-Graber. 2013. Topic models for translation domain adaptation. In *Topic Models: Computation, Application, and Evaluation. NIPS Workshop*.

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.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

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

Prashant Mathur, Marcello Federico, Selçuk Köprü, Sharam Khadivi, and Hassan Sawaf. 2015. Topic adaptation for machine translation of e-commerce content. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Tasnim Mohiuddin, Philipp Koehn, Vishrav Chaudhary, James Cross, Shruti Bhosale, and Shafiq Joty. 2022. Data selection curriculum for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1569–1582, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for java.

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

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134.

Maja Popović. 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.

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. 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.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70, Baltimore, Maryland, USA. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*.

Andy Way. 2013. Traditional and emerging use-cases for machine translation. *Proceedings of Translating and the Computer*, 35:12.

Deyi Xiong, Min Zhang, and Xing Wang. 2015. Topic-based coherence modeling for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):483–493.

Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*, pages 233–242. ACM.

Jian Zhang, Liangyou Li, Andy Way, and Qun Liu. 2016. Topic-informed neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1807–1817.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. *CoRR*, abs/1808.08266.