Leveraging Latent Topic Information to Improve Product Machine Translation

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

Abstract

The expectations of e-commerce customers include the ability to shop online in their preferred language. Modern e-commerce platforms utilize machine translation to provide multilingual product information at scale. However, maintaining machine translation quality that keeps up with an ever-expanding product information remains an open challenge for industrial machine translation systems. Topical clustering provides latent signals and interpretable textual patterns that can potentially help to improve translation quality and manage industry-scale translation data discovery. In this paper, we propose a topic-based data selection and a topic-signal augmentation methods using latent topic clusters to improve e-commerce machine translation quality. Additionally, we present a data discovery workflow using topic clusters to better manage expanding multilingual product catalogs.

Keywords: *product information translation, topic signal augmentation, topic-based data selection, textual pattern extraction, topical clustering, data discovery*

1 Introduction

With the advent of localized e-commerce sites, customers can now shop in their preferred language. Modern e-commerce platforms 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) with machine translated product information, titles, descriptions, and bullet points (Way, 2013; Guha and Heger, 2014; Zhou et al., 2018; Wang et al., 2021).

As e-commerce product catalogs expand over time, keeping machine translation systems up-to-date can be a challenging task. Product information is constantly sourced from different sellers that present their data stylistically. This lead to source data inconsistencies and translation inaccuracies, and validating this large amount of data for MT training at scale becomes an increasingly challenging and time-consuming task that requires substantial resources to manually review and correct errors in the training data to ensure accurate interpretation of product information.

From data usage perspective, topic words extracted from the latent topic clusters in the training data may be mapped to topics that can help with word sense disambiguation and improve overall performance of MT. Therefore, in this study, we propose two approaches to leverage latent topical clusters to improve machine translation quality: (i) **topic-based data selection** and (ii) **topic-signal augmentation**. Both approaches use Dirichlet Mixture Model (DMM)

(Nigam et al., 2000) with Collapsed Gibbs Sampling (CoGS) (Yin and Wang, 2014) to cluster large volumes of data efficiently, and the number of the clusters can be inferred automatically. The topic-based data selection approach first distinguishes between clusters of clean desirable data and those of noisy undesirable data based on the inspection of textual patterns, then selects training data from the desirable clusters for MT training. The topic-signal augmentation approach extends training data with extracted latent topic words prefixed to the source input and augments MT training with additional contextual information. Additionally, we propose a **data discovery workflow** to cluster training data and generates cluster summary and data visualization to uncover the latent topics and textual patterns, it can also identify new noisy data patterns so that strategies can be devised to prevent the occurrence of such data in the future.

2 Related Work

Previous researches 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) integrated 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 model to increase the likelihood of selecting words from the same topic as the source context. Instead of explicitly affecting the parameters or vocabulary selection, in this paper, we utilize a topic model for data selection and context augmentation implicitly adapting the model to the latent topic information.

2.1 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 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:

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

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:

$$P(w|z = k) = P(w|z = k, \Phi) = \phi_{k,w} \tag{2}$$

$$P(z = k) = P(z = k|\Theta) = \theta_k \tag{3}$$

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

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

z is the cluster labels of each document
m_z is 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

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 $P(Z_d = z|z_{\neg d}, d)$, $\neg d$ means d is not contained:

$$P(Z_d = z|z_{\neg d}, d) \propto \frac{m_{z,\neg d} + \alpha}{D - 1 + K\alpha} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{z,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{z,\neg d} + V\beta + i - 1)} \tag{4}$$

where, hyper-parameter α controls the popularity of the clusters, hyper-parameter β emphasizes on the similar words between a document and clusters.

3 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, we will further discuss the data discovery and inspection process in section 7. 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.

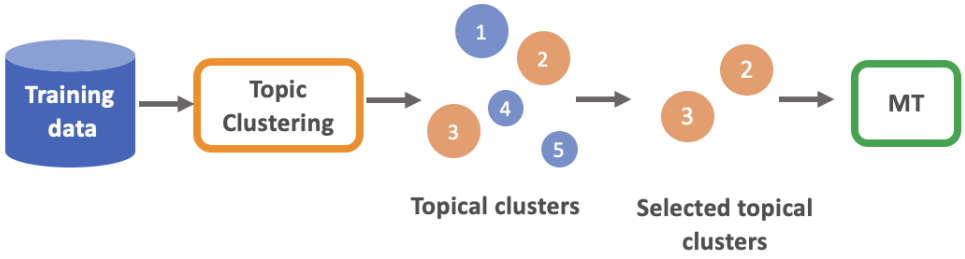

Figure 1: Choosing desirable data for MT training

4 Topic-signal augmentation

Figure 2 presents the topic-signal augmentation approach. We first cluster the data, then extract the most frequent top-k content words as the topic words for each cluster. Then, we choose the

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

larger clusters and prefix the source texts with the top-k topic words as topic signal as shown in Figure 4, we choose larger clusters are usually have more clear and interpret-able topic words, which can provide clear topic signals. Finally we extend the original training data with the training data with topic signal before training the MT model in Figure 3.

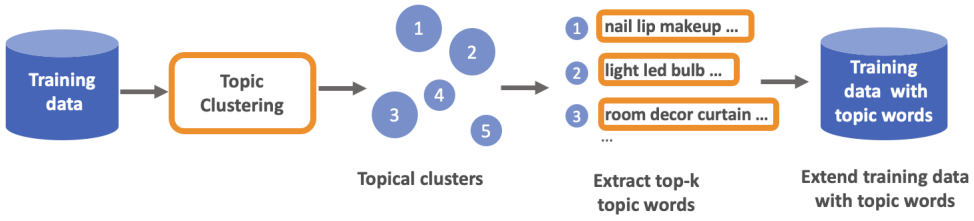

Figure 2: Topic signal approach to improve MT for product information

Figure 3: Use training data extended with topic words to further augment the MT training

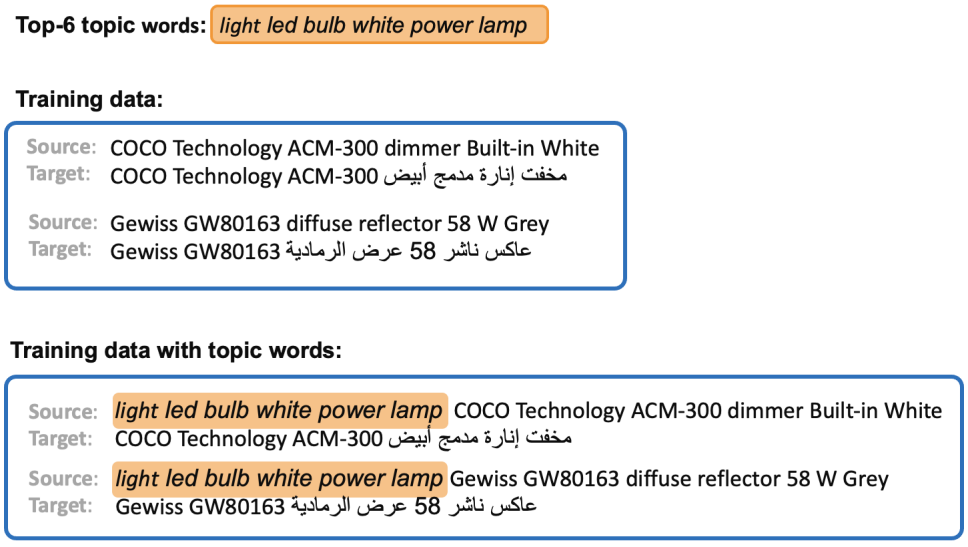

Figure 4: Extend the source text of the training data with top-k (k=6) topic words as topic signal

5 Experiment Setup

We experiment on four language pairs, English-Chinese (ENUS-ZHCN), English-Arabic (ENUS-ARAE), English-German (ENGB-DEDE), Spanish-English (ESES-ENGB). We train

the models on a large volume of in-house generic training data and a subset of product-information data (product titles, descriptions and bulletpoints) for domain adaptation. 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. For each language pair, we have three test data sets for product titles, descriptions and bulletpoints respectively. Each test data set has 2000 test segments and evaluate the models using BLEU[2] and chrF (Popović, 2015) to assess the translation quality.

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. For all 4 language pairs, the initial upper-bound number of topical clusters is set to 500 for ENUS-ZHCN and ENUS-ARAE, and 1000 for ENGB-DEDE and ESES-ENGB. 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 α and β 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 and inspect the topic words in the clusters created by the algorithm.

6 Experiment Results

6.1 Results: clustering results for four language pairs

	num of total clusters	num of major clusters	num of minor clusters	**% of data from the minor clusters**	**major cluster threshold**
ENUS-ZHCN	329	194	135	0.07%	50
ENUS-ARAE	374	110	264	1.32%	1000
ENGB-DEDE	536	117	419	0.09%	1000
ESES-ENGB	546	139	407	0.09%	1000

Table 1: Statistics of the Resulting Topic Clusters

Table 1 shows the statistics of the resulting clusters for each language pair. The number of the total clusters is automatically inferred by the algorithm. Each segments in the training data is assigned cluster IDs, and data selection uses a surprisingly simple heuristic-based human inspection of the data clusters.

We distinguish between major and minor clusters by the number of segments assigned to each cluster. For example, ENUS-ARAE clusters that contain more than 1000 segments are considered as major and selected as the training data for the model training; 1.32% of the training data with less than 1000 segments per cluster were dropped from the training data after the selection process. To yield a similar size data to the other three language pairs, we lower the major cluster threshold to 50 for the ENUS-ZHCN. The other language pairs have their major cluster threshold set at 1000. By removing minor clusters, 0.07%-1.32% data from the training data are removed.

Based on our inspection, we observe that the major clusters contain mostly the desirable data that captures the e-commerce themes, and the top-k words in the major clusters are intuitively good topic signals to improve machine translation. Meanwhile, we observe that the minor clusters capture undesirable various textual patterns that include noisy data. Therefore,

[2]SacreBLEU version 2.0.0 (Post, 2018)

we select only the data from the major clusters for our experiments. Section 7 will discuss further details on the data inspection process with an analysis for ENUS-ARAE translations.

6.2 Results: Improving MT with Topic Signals Augmentation and Topic-based Data Selection

As the described in Section 4, we extract the *top-6 most frequent content words*[3] from the major topic clusters and extend the source text to augment the original domain specific data to train the model. We refer to the models trained with augmented topic words as `Model Topi6`, and the models trained with the selected data from the major cluster as `Model Cluney`. We compare both `Model Topi6` and `Model Cluney` against the baseline models trained with the full in-domain dataset.

	Models	**Model Cluney**		**Model TOPI-6**	
	Domain	**BLEU**	**chrF**	**BLEU**	**chrF**
ENUS-ZHCN	Title	+1.40%	+2.87%	+1.77%	+3.56%
	Description	+0.58%	+1.63%	+2.85%	+4.12%
	Bulletpoints	+0.38%	+0.69%	+1.76%	+1.96%
ENUS-ARAE	Title	+7.20%	+2.79%	+3.03%	+0.09%
	Description	+1.45%	+0.87%	+2.14%	+0.32%
	Bulletpoints	+0.57%	+0.41%	+0.14%	0.00
ENGB-DEDE	Title	-0.11%	+1.23%	+1.70%	+0.55%
	Description	-0.43%	-0.17%	+0.80%	+0.46%
	Bulletpoints	+0.08%	+0.11%	+1.48%	+0.72%
ESES-ENGB	Title	+1.41%	+0.31%	+0.69%	+0.28%
	Description	-0.98%	-0.50%	+0.18%	-0.16%
	Bulletpoints	-0.58%	-0.31%	+0.51%	+0.04%

Table 2: Model Cluney and Model TOPI-6 Quality Improvement % over the Baseline Models

Table 2 presents the improved machine translation quality of the `Cluney` and `Topi6` model across language pairs and product information types. The `Topi6` model for ENUS-ZHCN reported the best improvements against the baseline with +2.85% BLEU and +4.12% chrF for description while the best `Cluney` improvements come from ENUS-ARAE with +7.20% BLEU and +2.79% chrF. The ESES-ENGB and ENGB-DEDE models have less improvement compared to the ENUS-ZHCN and ENUS-ARAE. It is possible that language pairs with similar source and target languages benefit less from the `Topi6` approaches.

Source	*100 GSM Comforter, Quality California **Queen** 400 Thread Count, 100% Egyptian Cotton*
Baseline MT	100 GSM 棉被,加州女王 400 支,100% 埃及棉
Topi6 MT	100 GSM 棉被,加州大号双人床 400 支,100% 埃及棉
Source	*Foot Fashion Lace Women's Dress Shoes, **Platform**, High Heel, Peep Toe (7.5, Rose Red)*
Baseline MT	Charm Foot 尚蕾女式正装鞋,高跟,露趾(7.5,玫瑰)
Topi6 MT	Charm Foot 尚蕾女式正装鞋,防水台,高跟鞋,露趾(7.5,玫瑰)

Table 3: Translation examples with improved word sense disambiguation

[3] $k = 6$ is chosen arbitrarily for this study, we will further investigate the impact of the span of the topic signal on the MT improvement in future work.

Anecdotally, we also observe some word sense disambiguation improvement in the translation especially in language pairs which the source and target languages are much different such as ENUS-ZHCN. Table 3 shows two translation examples of improved lexical disambiguation. In the first example, the word ***queen*** can refer to both *a person* and size semantically. In this case, it refers to the size of the comforter. The baseline MT incorrectly translates *the queen* to the person (女王) whereas the Topi6 MT successfully translates it to the size (大号双人床). In the second example, the word *platform* has a specific term in Mandarin when it refers to the platform of women's high-heel shoes. The baseline model conveniently ommits the translation for ***platform*** whereas Topi6 model translates it into the correct terminology 防水台.

7 Data discovery workflow and cluster inspection

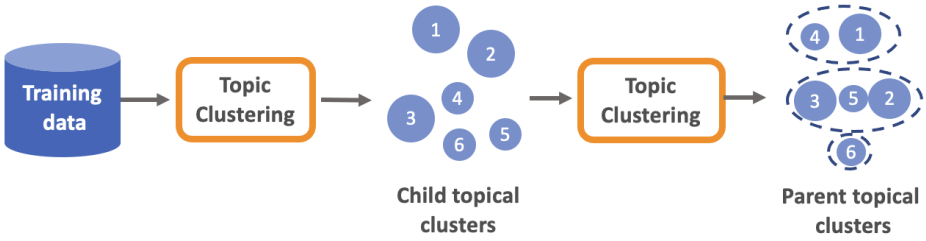

Figure 5: Data topic clustering workflow

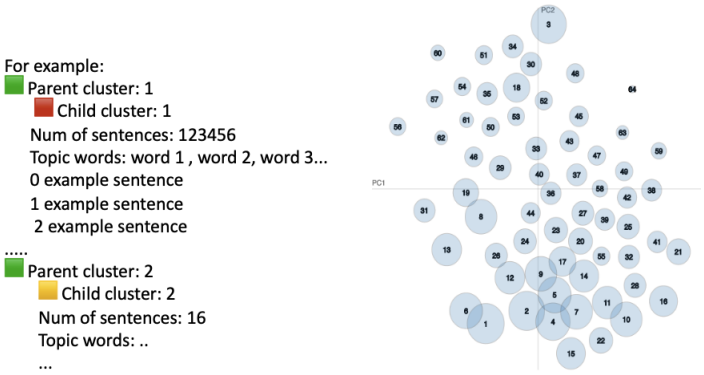

Figure 6: Left: Topic cluster summary for manual inspection. Right: 2-D plot for cluster visualization

Data quality management at scale for industry machine translation systems for an ever-expanding e-commerce product catalog is an open challenge. In this section, we describe a data discover workflow that clusters large volumes of data to allow human-computer interactive data inspection based on latent topic textual patterns. The process first cluster the texts into child clusters then optionally a second-stage clustering to create parent clusters as shown in Figure 6 on the left. Every cluster contains the segment IDs that fall within the cluster and a list of topic words that represent the textual patterns of the cluster.

This is particularly useful when acquiring training data from multiple sources on a regular basis or working with new language pairs, where unforeseen patterns may exist in the data. For each child cluster, we label its size with a color-coded square, where red is for large clusters

(e.g., $\geq$ 1K data points) and yellow is for smaller clusters. We sample a few sentences from each child cluster for manual inspection, along with the top-K most frequent content words to indicate the thematic information of the cluster. If child clusters are further clustered into parent clusters, we group them into a green square-labeled group cluster. We also generate 2-D data visualization with projected child clusters to understand the relations of clusters as Figure 6 on the right.These 2-D data plot can be generated using Jensen-Shannon distance (Fuglede and Topsoe, 2004) and multi-dimensional scaling technique such as Principle Coordinate Analysis (PCoA) (Borg and Groenen, 2005) .

We use ENUS-ARAE language pair to illustrate of our data inspection findings. For this language pair, there are total of 374 child clusters and 155 parent clusters returned from the data discovery workflow. Figure 7 displays the plots of all the 374 child clusters, where the size of the child clusters in the plot corresponds to the number of segments labeled for the clusters. Upon plotting all the 374 clusters, we observe a long tail of small clusters that deviated from the major clusters.

To gain further insight into the latent patterns of each cluster, we generate a cluster summary with sample sentences and parent clusters. We observe there are 110 major clusters, each containing $>=$ 1000 segments. The topic words and sample sentences in the cluster summary making it easy to infer the themes of these clusters. For instance, there are clusters having clear themes such as beauty and toiletry products or author names in European languages.

Meanwhile, many smaller clusters indicate undesirable textual patterns. For instance, there are clusters containing source texts in mixed English and Arabic, which also appears on the target side. Some clusters comprise source texts entirely in Arabic, with only some of the data appearing on the target side. There are also several clusters in which the source texts are in other languages besides the target side. Some clusters exhibit different noise patterns, which also appeared on the target side. Therefore, we use the major clusters in both experiments of the proposed MT improvement approaches.

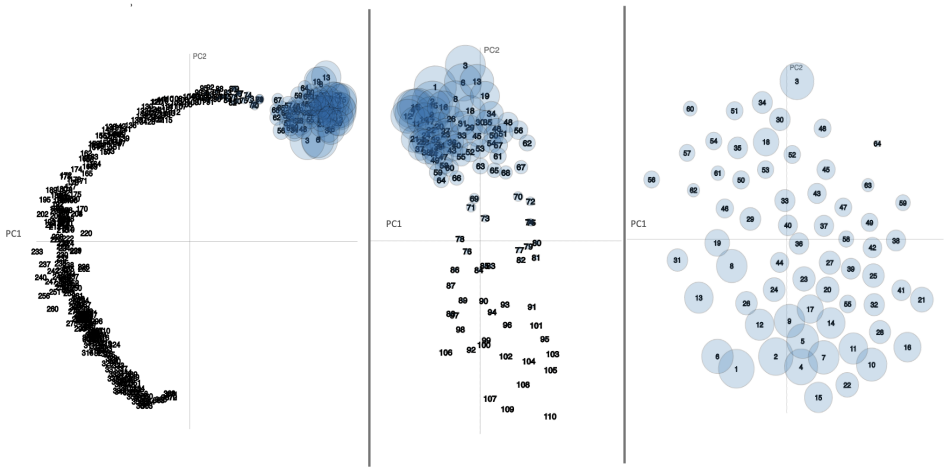

Figure 7: Data clusters of the English data (source text) from the ENUS-ARAE training Data using PCoA. The plot on the left is the visualization of all 374 clusters. The plot in the middle is for the top 110 clusters which size is $>=$ 1K sentences. The plot on the right is for the top 64 clusters which size is $>=$ 100 K sentences.

8 Conclusion

In this paper, we propose **topic-based data selection** and **topic-signal augmentation** approaches that leverages latent topic information to improve machine translation quality. Our experiments show that topic-based data selection and topic-signal augmentation works better on source and target languages that are more dissimilar (ENUS-ARAE and ENUS-ZHCN) than translations between similar languages (ENGB-DEDE and ESES-ENGB). Additionally, the latent topic words and clusters creates a data discovery workflow that allows manual data inspection and translation data quality control.

References

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.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Chen, K., Wang, R., Utiyama, M., Sumita, E., and Zhao, T. (2019). Neural machine translation with sentence-level topic context. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12):1970–1984.

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.

Eidelman, V., Boyd-Graber, J., and Resnik, P. (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, Jeju Island, Korea. Association for Computational Linguistics.

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

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

Hu, Y., Zhai, K., Edelman, V., and Boyd-Graber, J. (2013). Topic models for translation domain adaptation. In *Topic Models: Computation, Application, and Evaluation. NIPS Workshop*.

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.

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

Mathur, P., Federico, M., Köprü, S., Khadivi, S., and Sawaf, H. (2015). Topic adaptation for machine translation of e-commerce content. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

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

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134.

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.

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. (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.

Sievert, C. and Shirley, K. (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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Wang, H., Wu, H., He, Z., Huang, L., and Church, K. W. (2021). Progress in machine translation. *Engineering*.

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

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

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

Zhang, J., Li, L., Way, A., and Liu, Q. (2016). Topic-informed neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1807–1817, Osaka, Japan. The COLING 2016 Organizing Committee.

Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. *CoRR*, abs/1808.08266.