Machine Translation for Global E-Commerce on eBay

Jyoti Guha, Carmen Heger jyoti.guha@ebay.com, caheger@ebay.com
eBay Inc., 2065 Hamilton Avenue, San Jose, CA 95125, United States

Abstract

This paper describes the application of machine translation in overcoming language barriers for global e-commerce at eBay. We look at how buyers are provided the ability to perform cross-lingual search by translation of their search queries into the language of the listed items. We also show the process of translating item titles and descriptions from the inventory language into the user's preferred language. We briefly discuss the role of translation in facilitating communication between buyers and sellers.

In addition, we highlight the merits of evaluating translation quality together with system performance, which for search favors position-invariant metrics over more widely used translation metrics. Finally, we show how we incorporate user behavior and conversion metrics as part of our system evaluation methodology.

1. Introduction

With the global expansion of e-commerce, we are seeing an increase in cross-border trade amongst countries. At present, about 22% of eBay Inc.'s business is due to cross-border trade and about 30% of new customers are acquired via cross-border trade. This leads to situations where the buyer may not understand the language in which products are listed.

At eBay, we are using machine translation (MT) to help overcome this language barrier along the user's path, from discovering the item to making the purchase. Currently, we are supporting this use case for Russia, Brazil and Latin America in their respective languages, Russian, Portuguese and Spanish.

Search plays a pivotal role in the buyer's ability to find products. This makes accuracy and speed of the translation of a search query critical to enhance user experience and increase revenue.

Listing titles are seen throughout the conversion path on the site: in the "Search Results Page" which displays the list of results retrieved based on user search, the "View Item Page" which provides detailed information on the selected item, as well as the "Checkout Page" and other pages in the purchase path funnel. Correctness with respect to the context and consistency of these translations is essential to help users complete their purchases and finally achieve the best possible user experience.

In each of these areas, user-generated data, a broad array of domains, and the presence of domain-specific jargon present significant challenges for machine translation.

As an example, let us go over the use-case for Russia. A Russian user is provided with the ability to search in Russian. The user's query in his native language is machine-translated to English and this translated term is used to retrieve items with English titles and descriptions. Once the relevant items, originally listed in English, are fetched, their titles and descriptions are translated by default and displayed in Russian.

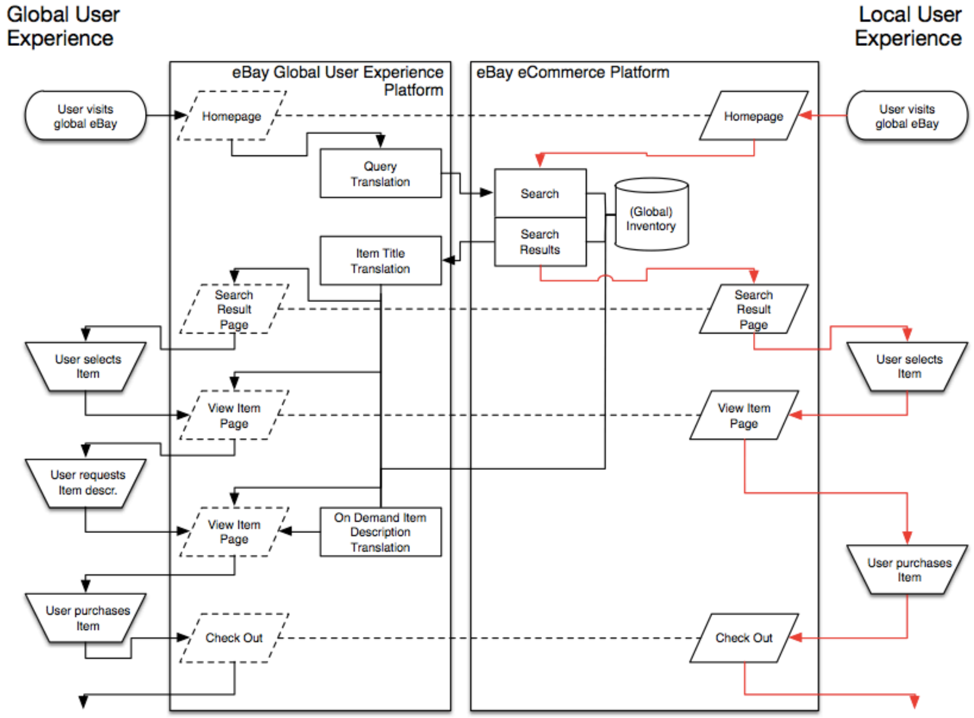


Figure 1: Excerpt of the transaction flow in a Global eCommerce Use Case, and the way of integration of the *eBay Global User Experience Platform* layer with the general purchasing workflow (the *eBay eCommerce Platform*). The schematics shows also a comparison of workflows of Global vs. Local User Experience (left vs. right user flow, respectively).

This helps the user understand the details of what the item is about and enables him to make a purchase decision, i.e., the item translation is "actionable". In addition, seller and buyer are provided with an inline translation option to facilitate communication amongst each other.

Our MT system treats each content type, namely search, title, description and communication, in the aforementioned flow as distinct use cases due to varying quality and performance needs. These needs give rise to unique sets of challenges for each of the content types; the subsequent section of this paper will cover these challenges.

We will also discuss how the system is tuned and evaluated for the specific use cases and look at the merits of evaluating translation quality together with system performance. Finally, we show how we incorporate user behavior and conversion metrics as part of our system evaluation methodology and in selecting training data as well.

2. Influencing Factors

Our overall goal is to provide eBay's international buyer the best user experience in their desired language, which is influenced by response time and quality.

2.1. Quality

Users can choose to enter search terms in the language of their comfort, which typically wavers between the user's local language and English. The translation system needs to preserve the English terms as-is and translate the local language to English in order to facilitate cross-lingual search. Given that the search terms on an average have four words, a relatively short length for language identification, discerning the source language can be challenging; especially for similar languages such as Portuguese and Spanish. This leads to a risk of unintended translations of English terms.

A high percentage of eBay titles have a 'non-standard' source language structure and style as they may use eBay-specific jargon and contain brand names, which should be preserved. The same brand names can also exist as regular nouns though, e.g. the word 'Gap' is a brand name but can also be treated as a regular noun based on the context. Also, a large amount of textual descriptions in eBay's inventory is produced by non-native speakers, often with mediocre experience in the respective language of the site on which the listing is hosted.

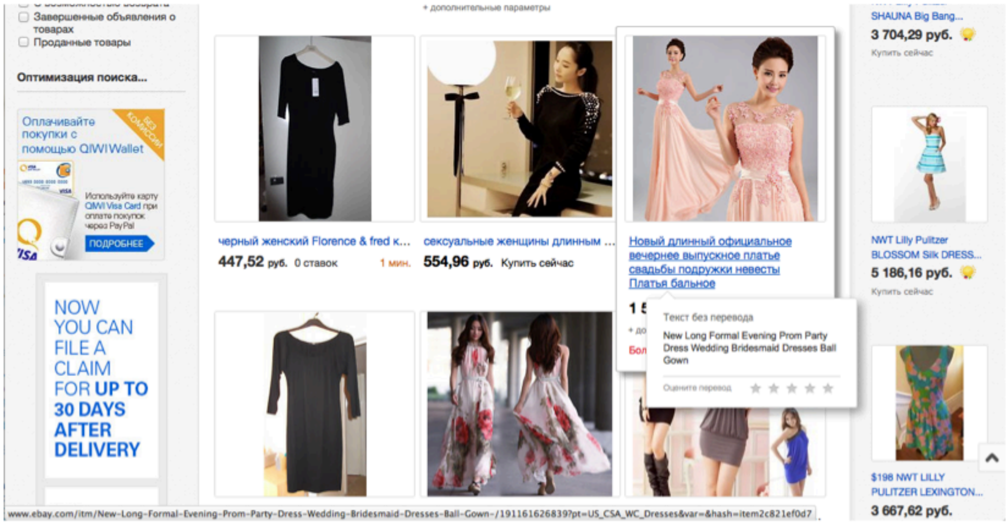
Coming up with a task-based quality evaluation in addition to standard quality metrics for each of the content types is non-trivial and we will go over the details in the upcoming sections.

2.2. User Experience

On the one hand, the search term translations are integrated in the real-time flow and must be seamless from the user perspective. Therefore low response times are mandatory and the translation has to balance quality and speed in order to achieve these latency requirements.

On the other hand, translated titles need to be displayed along with the option to see their original versions, hence it is important to come up with a user interface, which is non-intrusive but serves the purpose.

At present we translate roughly 820M characters per day across the 3 languages, Russian, Spanish and Portuguese. Processing such high volume within required latencies is also challenging for the overall translation system.



□ Завершенные объявления о товарах
□ Проданные товары
Оптимизация поиска...
+ дополнительные параметры
Оплачивайте покупки с помощью QIWI Wallet
БЕЗ КОМИССИИ
Используйте карту QIWI Visa Card при оплате покупок через PayPal
ПОДРОБНЕЕ
NOW YOU CAN FILE A CLAIM FOR UP TO 30 DAYS AFTER DELIVERY

черный женский Florence & fred к...
447,52 руб. 0 ставок 1 мин.

сексуальные женщины длинным ...
554,96 руб. Купить сейчас

Новый длинный официальное вечернее выпускное платье свадьбы подружки невесты Платья бальное

Текст без перевода
New Long Formal Evening Prom Party Dress Wedding Bridesmaid Dresses Ball Gown
Оцените перевод ☆ ☆ ☆ ☆ ☆

SHAUNA Big Bang...
3 704,29 руб.
Купить сейчас

NWT Lilly Pulitzer BLOSSOM Silk DRESS...
5 186,16 руб.
Купить сейчас

$198 NWT LILLY PULITZER LEXINGTON...
3 667,62 руб.

www.ebay.com/itm/New-Long-Formal-Evening-Prom-Party-Dress-Wedding-Bridesmaid-Dresses-Ball-Gown-/191161626839?pt=US_CSA_WC_Dresses&var=&hash=item2c821ef0d7

Figure 2: Search Result Page for the search term "платье", showing a Russian user's results with automatic translation into Russian. Original text can be seen, when hovering with the mouse over the translation. Also, the user is able to provide feedback to the MT system, allowing to incorporate this feedback to improve machine translation quality.

3. MT System

At a high level, the architecture consists of two layers: the MT Service Layer and the MT Engines Pool. The MT Service serves as the entry point and an orchestration layer. It is responsible to parse the input context (source/target language, context type: query, title and others) and choose the right underlying MT engines to serve the request. In addition, this layer performs necessary authentication, enables experimentation with different pipelines based on traffic allocation and logs any traffic information. Each MT engine in the pool specifies a pipeline built on language processing steps like tokenization, sentence detection, lowercasing, recasing, etc.

Currently, we use Moses, an open source framework for statistical machine translation, along with its XMLRPC service implementation (Koehn et al. 2007). The MT system is a standard phrase-based system trained on both eBay data as well as data from other domains (e.g. Europarl, CommonCrawl, corpora available through TAUS). While the first is crucial for quality the latter is important for vocabulary coverage.

For adapting the MT system to the content on the eBay site, as input by eBay users, we have generated bilingual data by translating search queries, item titles and item descriptions by human translators. Since these data sets are costly and time consuming to create, we can only select a comparatively small portion. Looking at user data we sample from what users frequently search for and what they often see. Also we try to select diverse content to cover several categories and styles. These corpora are used for both training and testing. In addition, we use monolingual eBay content for language modeling and evaluation. Finally, we include e-commerce-specific data like brands and acronym lists which help with translating those types more accurately.

4. Evaluation

The MT system is evaluated for both translation quality and impact. At eBay, we have the advantage to see what the user does with the MT output. We can measure the impact of MT by analyzing user behavioral data, e.g. a click on an item or a purchase, and thus can implicitly measure MT quality.

For the explicit evaluation of MT quality we look at a variety of automatic measures as well as human judgments.

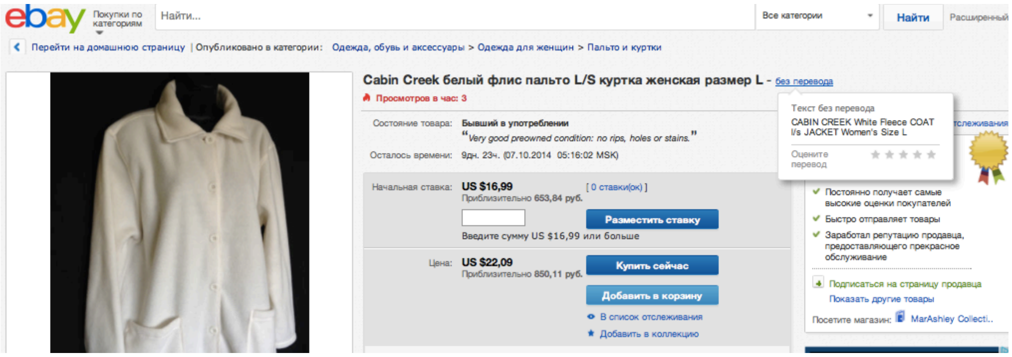
4.1. BLEU

BLEU is an n-gram based precision score that is commonly used for automatically measuring MT quality (Papineni et al., 2002) Most of our systems are optimized on BLEU and we use it, amongst other metrics, to compare different candidates.

4.2. PER

In some cases, we do not care about the order of the words in the translation. For instance, the query "*used musical instruments classical guitar*" leads to the same search results as "*musical instruments classical guitar used*".

BLEU penalizes the "wrong" word order, which results in 70.71% for the given example. PER, here 0%, is a position-independent word error rate, which in this case relates much better to the quality of the translation in its given use case.



Опубликовано в категории: Одежда, обувь и аксессуары > Одежда для женщин > Пальто и куртки

Cabin Creek белый флис пальто L/S куртка женская размер L - без перевода

Просмотров в час: 3

Текст без перевода
CABIN CREEK White Fleece COAT l/s JACKET Women's Size L
Оцените перевод ☆☆☆☆☆

Состояние товара: Бывший в употреблении
"Very good preowned condition: no rips, holes or stains."

Осталось времени: 9дн. 23ч. (07.10.2014 05:16:02 MSK)

Начальная ставка: US $16,99 [0 ставки(ок)]
Приблизительно 653,84 руб.
Введите сумму US $16,99 или больше

Цена: US $22,09
Приблизительно 850,11 руб.

✓ Постоянно получает самые высокие оценки покупателей
✓ Быстро отправляет товары
✓ Заработал репутацию продавца, предоставляющего прекрасное обслуживание

Посетите магазин: MarAshley Collecti..

Figure 3: View Item Page showing the title that has been automatically translated to Russian. Original text can be seen, when hovering with the mouse over the translation. The user is able to provide feedback to the MT system; this feedback is incorporated to improve machine translation quality.

4.3. Brand Preservation

E-commerce data consists of many brand and product names. Often those names are or contain words that are also part of a language. E.g. the brand "REI" means "king" in Portuguese and is likely to be translated as such by a Portuguese to English MT system that is trained on lower-cased data.

While the context should usually guide the translation process to the correct translation, it is especially hard for search queries because the context is extremely small and often even just the word by itself.

The correct translation of brand and product names is crucial for search queries, item titles and descriptions, which is why we maintain a list of brands and products and feed it to the MT system.

When evaluating the translation of brands we look at the preservation rate by counting how often those brands are kept as-is and not translated. Our MT systems preserve >90% of the brands.

4.4. Search Results

Correct translation does not always mean good search results. E.g. the Russian query "*сумки из натуральной кожи*" can be translated into "*bags of genuine leather*" or "*genuine leather bags*". Both are valid translations but the latter one leads to many more search results (47,000 vs. 160).

This is why we evaluate search recall for query translation, i.e. we look at the number of queries for which we get no results and compare the average result set sizes. The results are weighted by query frequency in order to get a more realistic distribution.

Only in <10% our MT systems produce queries with null results. Please note that we currently only look at the pure recall, i.e. number of search results, and not at their relevance.

4.5. Human Judgement

Evaluation by human judgments is done in 2 different ways:

- **Translation quality**: This metric is rated on a 1-5 scale (for titles) or simply good/bad (for queries). A bilingual speaker rates the quality of MT output given a source string;
- **Understandability**: This metric is also rated on a 1-5 scale but aims at evaluating how well a translated title is understood without knowing the source language. A monolingual speaker subjectively rates the quality of MT output given only the translated string and the picture of the item.

All our currently deployed MT systems reach a quality that our users consider *understandable* and *actionable*.

5. Impact and Results

We have seen a positive impact of machine translations based on some of the key user behavioral metrics measured after the feature launch, as well as direct user feedback.

5.1. Results in Search

Adoption of search in the users' local language increased by 18%. This number indicates the buyers' confidence in the translation and signals a positive influence when providing them a comfortable and efficient shopping experience.

Using our in-house translation system has helped meet the aggressive site latencies, decreasing unsuccessful item queries by 45%.

5.2. Item Title Translations

We surveyed a large group of users to gather feedback on their experience with our automated title translation, resulting in the following readout:

- 78.3% of users felt that their shopping experience on eBay increased;
- 82.3% rated the machine translation acceptable to high quality (i.e. MT results are actionable);
- 82.5% would recommend the automatic translation of all item titles on eBay

6. Conclusion

We presented eBay's integration of machine translation into the e-commerce workflow to enable global customers find the products they desire, across country and language borders. We also looked at the influence of the commerce-specific machine translation on the user experience.

While we will continue to augment the language pairs supported by machine translation and improve the machine translation quality of the currently deployed languages, we will expand the machine translation to other areas and content types on the site to enhance the user experience even further.

Acknowledgments

We want to thank all the eBay MT team and localization team members for all their contributions and feedback.

MT team: Derek Barnes, Irina Borisova, Nishit Chokhawala, Steven Cook, Saurabh Dhupar, Mithun Gaddam, Braddock Gaskill, Ethan Hart, Saša Hasan, Mahesh Joshi, Saiyam Kohli, Selçuk Köprü, Michael Kozielski, Gregor Leusch, Asim Mathur, Evgeny Matusov, Kiran Nagarur, Jean-David Ruvini, Sumita Sami, Hassan Sawaf, Yoram Vardi, Mirko Vogel, Mudar Yaghi.
Localization team members working with the MT team: Tanya Badeka, Marta Choroco, Paula Foster, Tatiana Kontsevich, Silvio Picinini, Olga Pospoleva, Juan Rowda, Jose Sánchez, Jen Wang.

References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, PA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation,* Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.