Challenges of Machine Translation for User Generated Content - Queries from Brazilian users

Silvio Picinini spicinini@ebay.com

Abstract

Users of eBay in Brazil enter their queries in their native language, Brazilian Portuguese. Those queries must be translated into English to search the vast inventory of products available on Ebay. Given the volume and variety of queries, typical for user generated content, make this scenario a prime use case for Machine Translation. This presentation shows the findings in user generated queries that are challenging for the MT engine to translate correctly.

1. Introduction

Users of eBay in Brazil enter their queries (search terms) in their native language, Brazilian Portuguese. Those queries must be translated into English to search the vast inventory of products available in English on Ebay.com (the company's US site). The volume and variety of the queries make this scenario a prime use case for Machine Translation (MT), and a new territory to explore. This presentation shows the findings in user generated queries that present challenges for the MT engine.

General challenges of user-generated content

The type of user-generated content addressed by this presentation has certain features. Some may be common to many other types of user generated content, while others may be specific to search queries created by users.

1. Large content volume

A high number of users create a volume of content (around 1 M queries daily) whose translation can only be performed in an automated way, with MT.

2. Widely varied content

A large number of users create a broad range of interests from these users, because each one has a different interest. Therefore, the content becomes highly diverse, covering a wide range of subjects, product categories and product features.

3. Unstructured and incomplete content

A query could be written as "*I am looking for a cap with the logo of the Raiders football team*". It would make the work of the MT engine easier because this is a complete sentence with all the structure and all the parts of a sentence. The query above could be a query that one would ask a human person at a store. But a query online is different. Users normally enter only keywords. Therefore, a query could be just "cap". This query lacks clarity and is ambiguous. This makes it more difficult for the MT to be accurate because search queries are unstructured and are not a complete sentence.

4. Lack of context

The query "*I am looking for a hard case to protect my phone*" could appear just as "hard case". The users know what their intent is. The MT, however, has the challenge of deciding whether "hard case" means a "rigid enclosure" or a "difficult court case". The MT engine has a decision to make, and it depends on the corpus that is being used to teach the engine. If the corpus is coming, for example, from European Parliament proceedings, which may contain regulatory and legal content, the engine could lean towards a "a difficult court case". This would be an incorrect translation.

5. Non-normalized content

The corpus that is used to train the engines usually comes from sources where the writing of the source language and the translation into the target language are done by professionals. This establishes a certain level of correctness of spelling and grammar as is expected from these professionals. It also establishes a certain formal tone for the language, consistent with the expectations of the target audience of that content. Last, the corpus may use terminology that is specific to the subject, a jargon. On the other hand, queries use everyday language. The spelling, grammar, and language of the queries depend on the user skills. With this language scenario in mind for the corpus and the queries, this topic is about "*how well the language of the queries matches the language of the corpus*".

The spelling and grammar of the queries may not match the spelling and grammar of the corpus. Users may write queries in a hurry and create disagreements in number (plural vs. singular) or gender (masculine vs. feminine), for example. This is rarely an issue in English, but the queries are coming, for example, from Romance languages where these issues are explicit. The queries may come from variations inside the language, such as different spellings in different countries. It is the equivalent of differences between English (US) and English (UK). The users may also just use common misspellings of words. Examples in English could be "their vs. there" or "affect vs. effect".

On the language tone, queries may be informal while the corpus is formal. In English, examples of informal constructions would be "4 u" or "2 u", instead of "for you" and "to you".

On language terminology, a more technical corpus could use "mobile device" to refer to a phone. The translation of a query such as "Barbie phone case" could become "Barbie mobile device enclosure", changing the tone and the search results.

Specific challenges in Brazilian Portuguese

Some of the challenges are situations that may occur only in Brazilian Portuguese, while some may appear in other languages.

One common pattern among the challenges is the use of everyday language. That kind of language may not be easily found among the sources of bilingual corpus available, because the corpus is usually written in a more technical or formal style, or is more precise in grammar and spelling for being professionally written.

It is worth noting that the examples in the topics below are all from real queries entered by eBay users.

1. Diminutives

Diminutives are used in Brazilian Portuguese to refer to things or people to express affection or cuteness. This also happens in other languages. The phenomenon is called Hypocorism and defined as:

> *A hypocorism is a shorter or diminutive form of a word or given name, for example, when used in more intimate situations as a nickname or term of endearment.*[1, 2]

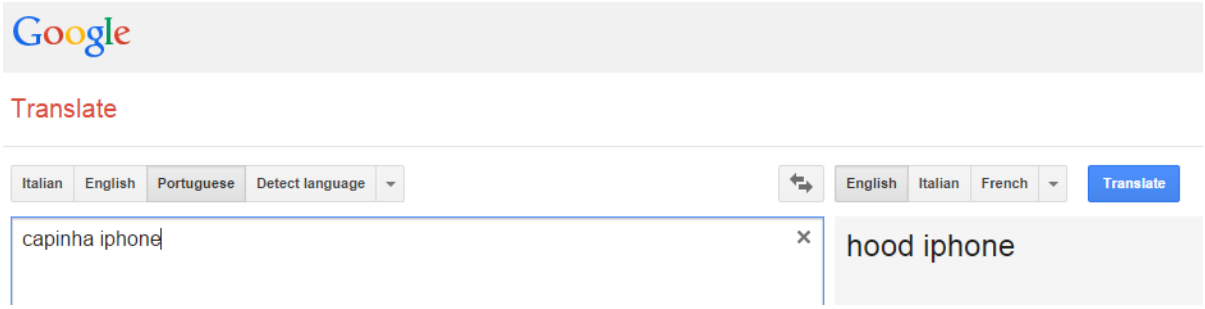
The equivalent example in English is to add a "y" to the name of a person (Vicky, Johnny, Joey). In English this occurs mostly for names of persons and for doggy, kitty and ducky, for example.

In English, the use of these words is concentrated on what is called "baby talk", accompanied by high pitched voice in many cases. However, in Brazilian Portuguese, hypocorism is widespread. It can be used for all kinds of objects, such as a case or a blouse. Virtually anything that could use the adjective "tiny" in English could fit as a diminutive in Brazilian Portuguese. The diminutive is incorporated into the words as a suffix, usually "inho" or "inha". For the "case" of a cell phone the word "capa" becomes the diminutive "capinha". The English

equivalent would be to call a cute little case as a "casey", a doll house as a "housey" and so on.

The challenge for MT is to translate words that are not usually part of the available parallel corpus used to train the engine. In English, the equivalent is the translation for "doggy". If the source of corpus is the European Parliament, or even veterinary material from a University, the content is formal or technical, and we may find "dog" but not find "doggy". This affects the machine translation of "doggy".

In Brazilian Portuguese, the use of diminutives is more widespread, so the challenge to translation is also more frequent. One example of the consequences of that is shown below. The translation for "capinha" (little case in Brazilian Portuguese) appeared on Google Translate as "hood". The correct translation would be "case".



Examples of diminutives:

- Capinha (little case)
- Carrinho (little toy car)
- Varinha (little wand)
- Botinha (little boots)
- Cadeirinha (little chair)
- Bichinho (little animal)
- Blusinha (little blouse)
- Ursinho (little bear)
- Roupinha (little clothing)

Frequency

A set of **over 2 million** queries was used for analysis. The most common diminutive term (capinha) alone appeared in **1 in every 1000** queries. Next, we

looked at the 31 most common diminutive terms. These terms appeared **1 in every 500** queries. These numbers clearly indicate that diminutives are a relevant issue for PTBR.

Solution

The solution for this challenge is **to obtain "in-domain" data** containing the style used by people when writing these queries. Training the engine with correct data will improve the MT output.

We obtain in-domain data through the translation and post-editing of eBay content containing queries, item titles and item descriptions.

The proof that this solution is effective is the fact that some of the diminutives that challenge the MT engine are already being correctly translated. The more we obtain or create data to train the engine, the better diminutives will be translated.

2. Lack of diacritical marks

Accents: This is barely a problem for English, because the language rarely has accents. The few words that have accents are mostly originated from other languages. One example is résumé, which can even be spelled with two accents. But the accents exist in other languages, such as Brazilian Portuguese.

When entering a query, the user may not use accents. The challenges for the MT engines are:

- The MT engine should learn all variations of a word, with or without accents.
- The corpus, if coming from professionally written sources, will tend to be correct and will contain accents. Therefore it may be difficult for the MT engine to learn the "incorrect" variations that do not contain the accents.

Examples in Brazilian Portuguese:

- relogio vs relógio (watch)
- oculos vs. óculos (glasses)
- camera vs. câmera (camera)
- maquina vs. máquina (machine, camera)
- tenis vs. tênis (sneakers)

Other diacritical marks: Portuguese contains the characters **ç**, **ã**, **õ**. These are often spelled as the regular letters c, a and o. The challenges for MT are:

- The MT engine should learn all variations of a word, with or without the diacritical marks.
- The corpus, if coming from professionally written sources, will tend to be correct and will contain the diacritical marks. Therefore it may be difficult for the MT engine to learn the "incorrect" variations that do not contain the accents.

Examples in Brazilian Portuguese:

- coração vs coracao (heart)
- macacao vs. macacão (romper)
- violao vs. violão (acoustic guitar)
- calca vs. calça (pants)
- promoçoes vs. promoções (promotions)

Frequency

For accents, relógio appeared correctly with the accent 21% of the time and without the accent 79% of the time. This word alone appeared in **1 in 100** queries.

From the set of queries, 88% contained calça with ç and 12% contained calca. Cartão with ã appeared in 46% of the queries and cartao appeared in 54%. Macacão appeared in 66% and macacao in 34%. Peça (part) appears correctly with ç in 76% of the queries. Replacing the letter ç with the letter c it becomes a different word "peca" (meaning "he/she sins") and appears in 24%. These examples appeared in **1 in 400** queries.

This indicates that the issue is very frequent.

Solution

For accents as well as for the other diacritical marks, the solutions under consideration would be:

- **to obtain "in-domain" data** containing all the variations of a word – this is already in place and some of the words are already being correctly translated

- **normalization**
    - normalize all data to no accents – however this would introduce ambiguity because two different words (one with the accent and another without it) with two different meanings would become the same word
    - normalize all data to no diacritical marks – less risk of ambiguity
- **spelling** correction

All of the examples above are being correctly translated, which indicates that having in-domain data is an effective solution.

3. Words intended for the target language that happen to be also words in the source language.

In some cases, the query from the Brazilian user contains a word that is intended to be understood in English, but that word is also a valid word in Brazilian Portuguese. The challenge for MT is to identify that intention. This is an ambiguity issue and could be seen as a "polysemy across languages", with the same word having two different meanings in two different languages.

Ideally the engine would leave the word untranslated, since the intention was already an English word or brand. If this does not happen, the MT engine will consider the word as a Brazilian Portuguese word and translate it into English with a completely different meaning.

Examples in Brazilian Portuguese:

Costumes - Costumes translated as the PT word "costumes" which means "customs". The real intention was to search for the EN word costumes.

Fins- Translated as the PT word "fins" which means "purposes". The real intention was to search for the surfing accessory called fins.

Cruze - "Cruze" was translated as the PT word "cruze" which means "to cross". The search is looking for the car Chevy Cruze.

Frequency

The word "costume" was frequent in the set of queries. From these, 3**%** had an incorrect translation. For Cruze, 100% of the queries had incorrect translations. Fins had 34% incorrect translations. These issues appeared in **1 in 600** queries.

Solution

One solution has been to resolve very frequent issues on case-by-case basis. Another solution is to submit both the translated query and the original query to the search engine and obtain results from both. These results could be mixed. Other solutions could involve language identification from the context or from the history of queries by the user. These are more complex and the benefit may not justify.

The issues below are **not among the top most frequent** and were found in other query sets that were analyzed.

4. Homophones

Some words sound the same when pronounced. They are called homophones, and one definition is:

> *(Linguistics) one of a group of words pronounced in the same way but differing in meaning or spelling or both, as for example bear and bare. (Collins Dictionary)*[3]

These words may be misspelled in a query and become a challenge for the MT. English examples would be "knight" and "night", or "your" and "you're".

In Brazilian Portuguese some letters or groups of letters have the same sound. The users may misspell a word using a letter with the same sound. This creates homophones:

- The letters s, z and x may have the same sound.
    - Digitalizadora spelled as digitalisadora
- The letter groups c, ç, ss, sc, xc and in some cases s may have the same sound.
    - Engraçados spelled as engrassados, acessórios spelled as asesorios, tensão spelled as tenção
- The letters ch and x may also have the same sound.
    - Puxadores spelled as puchadores, bichinho spelled as bixinho
- The letters g and j may also have the same sound.

This is a challenge for the MT engines because these words are less likely to be present in the training data. On the other hand, these issues are not of high frequency.

Solution

Some of these issues are covered by existing spelling correction mechanisms. For the issues that are not, the solution for this challenge is **to develop a correction model for homophones** that would consider the possibility of these same sounding words, enhancing the spelling correction. This is a cleaner solution than having in-domain data to cover wrong spellings.

5. Transcription

One definition for transcription is:

> *A written, printed, or typed copy of words that have been spoken*[4]

In this work, we are taking the liberty of expanding the definition of transcription to represent certain issues that we found. We are considering Transcription as "*the writing of words in the way they sound when they are spoken*".

In English, an example would be the word "dessert". Based on how it sounds, it could be spelled as "desert" (as in the verb to desert), or even "dezert". The musical instrument "bass" could be spelled as "base" or even "bace". This could be considered a form of transcription.

This creates a challenge for MT engine. In some cases, the intended meaning is changed. If the intention was "dessert" and the writing says "desert", the MT engine will not be aware of that. In the case of writing "dezert", this new word may not be present in the corpus and the MT will not provide a translation for it.

In Brazilian Portuguese:

- Words ending in eira spelled as era, or oura spelled as ora
    - Geladera instead of geladeira, tesora instead of tesoura
- Words with e spelled with i:
    - Vídeo spelled as vídio, folheada as foliada, pediatra as pidiatra
- Single consonants that are pronounced as if they had an accompanying vowel
    - p > pi (helicóptero as helicopitero, opção as opição)
    - t > ti (étnica for étinica)
- The vowel group ou spelled as o
    - roupa spelled as ropa
- The initial h is silent in many words
    - Helicóptero spelled as elicoptero

Solution

Similarly to the homophones, the solution for this challenge is to develop a correction model for transcriptions, enhancing the spelling correction.

6. Transliteration

One definition of transliteration is:

> *To represent (letters or words) in the corresponding characters of another alphabet.*

In this work, we are taking the liberty of expanding the definition of transliteration to represent certain issues that we found. We are calling transliteration "the action of a user trying to write in the target language using phonetic rules of the source language". It is not a different script or alphabet, but it is a crossing from one language into another, so we took the liberty of calling it a kind of transliteration.

The users in Brazilian Portuguese will, in some cases, write English words using the Portuguese constructions for the word. Examples:

- Maico Jordan instead of Michael Jordan
- Naik or Naiki instead of Nike
- bets instead of baths
- bim instead of beam
- bleide instead of blade
- yang instead of young

Solution

Similarly to the homophones, the solution for this challenge is to develop a correction model for transcriptions, enhancing the spelling correction.

Conclusion

These linguistic findings in Brazilian Portuguese show a variety of challenges and solutions for MT issues. Many issues **are frequent and require a more immediate attention**:

- Diminutives
- Lack of diacritical marks
- Words that are intended to be in English

Other issues are **not among the most frequent**, but once a solution is developed, the engine that has that solution will provide a better experience for a number of customers:

- Homophones
- Transcription
- Transliteration

Regarding the nature of the solutions, many issues are resolved by **having more in-domain training data,** so that the MT engine can learn from it. Our challenge remains in obtaining or creating this data.

Other solutions include the development of mechanism that will enhance the preprocessing of MT data and contribute to MT performance.

We will continue to work on improving the MT capabilities of eBay so that we can create a better translation and provide better search results and a better experience for our users.

Acknowledgements

I would like to express my deepest appreciation for the guidance and support received from Carmen Heger and Mike Dillinger.

References:

1. http://en.wikipedia.org/wiki/Hypocorism
2. http://grammar.about.com/od/d/g/diminterm.htm
3. http://www.thefreedictionary.com/homophone
4. http://www.merriam-webster.com/dictionary/transcription