Harvesting Polysemous Terms from e-commerce Data to Enhance QA

Silvio Picinini spicinini@ebay.com
Localization, eBay Inc., San Jose, CA, USA

Abstract

Polysemous words can be difficult to translate and can affect the quality of Machine Translation (MT) output. Once the MT quality is affected, it has a direct impact on post-editing and on human-assisted machine translation. The presence of these terms increases the risk of errors. We think that these important words can be used to improve and to measure quality of translations. We present three methods for finding these words from e-commerce data, based on Named Entity Recognition, Part of Speech and Search Queries.

1. Introduction

Polysemous are words or sets of words with multiple meanings. For this work, we consider a broader definition that reflects what polysemy causes in Machine Translation (MT). A polysemous word here is "a word that can have different translations". This means that the MT engine can be confused about which translation is the correct one; this in turn affects the quality of the machine translation. Instead of "polysemous", we could call these "polytranslation" terms, and just invent this word.

For example, a word that can be assigned multiple POS (parts-of-speech) tags may have a similar meaning, but it will be translated differently if it is a noun, a verb, or an adjective. Example: Print a report (verb), print magazine (adjective), this fabric has a nice print (noun).

A brand that is also a common word (e.g. Gap, Guess, Coach) will be left untranslated when referring to a brand and will be translated when used as a common word.

Also, a word like "mixer" may have a generic meaning of "a device that mixes", but in the real world, it can refer to very different products, such as a kitchen mixer or a sound mixer for music. These are two very different devices with very different translations. Also, it can be a party (singles mixer), a very different meaning.

This work presents three new processes that leverage eBay e-commerce data to harvest polysemous words, so that these can be used for different applications, such as the ones described in this paper.

Before going into the methods, we make two general points about data below.

2. Leveraging semantic value and relevance added by the public

We think that it is important to make a point about the importance of capturing semantic meaning from user behavior. It is a massive and no-cost source of information, therefore, we should be interested in using it. One of the methods described uses information from buyer behavior on eBay. By entering a query and then going into a certain category, the buyer is associating meaning to the query, which is comprised of one or two words. The same happens when a seller describes an item for sale and chooses a category for it, giving the words of the

item additional context and meaning. It is also important to capture relevance from user behavior. If we capture how frequent certain meanings are, we are capturing how relevant they are. This "quantification" is something that traditional dictionaries cannot do, and is a significant difference compared to dictionaries.

All of this can be seen as a "public semantic annotation" work that is being done without cost for the companies. While companies collect vast amounts of data, we are all constantly looking for ways to enhance the meaning of this data, and the examples in this work are in line with that overall effort.

3. Harvesting relevant words in context

Another important point about polysemous words and synonyms is these words are relevant because of their meaning relationship; polysemous words have a single form with multiple meanings and synonyms have multiple forms with a single meaning. There are also other types of relationships that can be of interest, such as hyponyms, hypernyms, and meronyms (a word that is a constituent part or a member of another word).

These words are useful in many ways, such as improving and measuring MT, but also improve search queries and possibly classification. The challenge is to find "applied" examples of these words in a specific context. While a dictionary or WordNet can tell us that words are synonyms, it will not tell us that "camcorder" is a synonym for "video camera" or that "flash drive" is the same as a "pendrive". Finding these words applied to a context, an industry, or a subject matter should be more useful than generic words.

4. Applications of polysemous words

There are some possible applications for polysemous words in machine translation:

- Select data containing these words and create training and testing data for them

Since these words are more likely to be mistranslated, they are more likely to require more in-context training data to help the engine disambiguate different situations. So one application of polysemous words is to find examples of content with these words and have it translated/post-edited. This will allow the creation of training data for the engine to learn how to better handle these words, and the creation of testing data to evaluate the translation.

- Evaluate the quality of the MT of these words

The evaluation of the MT output quality is usually performed on the entire content (using automated metrics) or on a sample (if human evaluation is used). In both cases, there is usually no "selection" of certain segments matching some certain criteria which should be measured, the segments are randomly chosen. However, polysemous words could be used to provide an insight on the quality of the machine translation, using selected "more difficult" words. eBay has started collecting some data around this.

- Evaluate the quality of the post-editing of these words

Training and testing data may be created through post-editing. The quality of that post-editing work needs to be evaluated. Polysemous words are more likely to be mistranslated and

be wrong in the MT output. The post-editing process is supposed to correct those errors. If the error is corrected by the post-editor, this is an indication that the post-editing is of good quality. If the error is not corrected, this is an indication that the post-editing may not be of good quality, and may need further work before being used as training or testing data. The evaluation of post-editing is usually done by an evaluator on a random sample.

Looking at how polysemous words were post-edited is a way to assess the quality of the post-editing work and is also an indication of the final quality of the content that is going to become training or testing data.

5. Three processes to harvest polysemous words

5.1. From eBay search queries

This process is based on associating different categories to the same query. The premise is that if a word is associated with two very different categories, they are likely to have very different meanings, and there is a good chance that the word is polysemous.

Customers enter search queries on eBay. After seeing the results of their queries, they take an action that leads to a certain category. This is an indication of the meaning of the word that was entered as a query. Let's consider an example with the word "mixer". A query for that word does not clarify if the customer is looking for a sound mixer or a kitchen mixer. However, after the query display results, the customer takes action to look into one of these different devices. Once the customer acts, there is now a category that can be attached to the word in the query.

eBay creates a column called Leaf Category Histogram. It looks like Figure 1:

15709/69.108|**6224**/14.247|**95672**/6.274|**6158**/1.696

Figure 1. Leaf Category Histogram

This column contains the identification of the most frequent categories (in black above) accessed by the customer after entering a query. It also contains the % of instances where the customer went to a certain category (in red above). This number is an indication of the "intensity of the polysemy". If a word like "mixer" goes 60% of the time to a music category (for sound mixer) and 40% to a kitchen category, this is an indication that there is significant interest for both meanings. If another word has a 99.8% frequency and the second category is, for example, less than 0.1%, this is an indication that one meaning is nearly universal and the other is extremely rare. This can inform our harvesting of polysemous words.

Starting from that data, we find the higher level eBay categories associated with the word. We are interested in finding big differences in categories, which would be more likely to have different meanings. Once we manipulate the data, we arrive at information that looks like Table 1:

Word	**Category 1**	**Frequency 1**	**Category 2**	**Frequency 2**	**Is Cat 1 diff from 2?**	**Is Freq 2 > 2%?**
Mixer	Music	60%	Kitchen	40%	Yes	Yes

Table 1. Data after manipulation

The last two columns are formulas. With this data we can filter the last two columns and the result will be a list of polysemous candidates. Table 2 show some examples we found in our initial results:

Word	Category 1	Category 2	Comment
Vans	Clothing, Shoes & Accessories	eBay Motors	brand vs. type of car
notebook	Computers/Tablets & Networking	Books	computer vs. writing
Fossil	Jewelry & Watches	Collectibles	brand vs. actual fossil
mixer	Musical Instruments & Gear	Home & Garden	sound table vs. dough mixer
roadrunner	eBay Motors	Toys & Hobbies	car vs. character
Pebble	Cell Phones & Accessories	Pet Supplies	brand of watches
torch	Sporting Goods	Business & Industrial	flashlight vs. hot flame

Table 2. Results from queries

A quick human triage to validate which of these candidates are good produces our final list.

The initial results indicate that this process is efficient. A list of about 1900 queries yielded about 40 candidates. A human triage that took about an hour yielded 19 final terms, about 1% of the initial data.

5.2. From NER data

For Named Entity Recognition, we tag individual tokens, mapping them to different tags according to their meaning. The premise for finding polysemous words in this process is that the same word can be tagged with different tags, and if these tags indicate a significantly different meaning, there is a good chance that the word is polysemous.

This process leans on the concept of polysemous words being defined by "how words are translated". The most benefit from this process comes from differentiating words that are not translated from words that are translated. The MT engine may be confused and translate brand names, or do not translate common words because they are commonly brand names. The word "charger" can refer to the car Dodge Charger. This is a product name and won't be translated. But it can also refer to a charger for a cell phone. This is a common word and will be translated. Therefore, it is possible that there is "a charger in a Charger", and the MT has to deal with this ambiguity.

We start with a list of tokens and tags for a certain category. Once we sort it by token, we will see that some tokens are tagged with different tags. Some NER tags indicate that the token should not be translated: Brand and Product Name. Other tags indicate that the meaning tends to be a common word: Type, Color. We organize the data with additional columns: Do Not Translate indicates when a token is tagged with Brand or Product Name. Translatable indicates when the token is tagged with a category that is usually a common word, and therefore translatable. Once the data is organized in this way, a few manipulations with sorting, filtering and formulas will produce the list of candidates that we are looking for.

Table 3 shows what the data looks like:

Word	Token	Do Not Translate?	Translatable?	Contains Translatable and DNT?
Charger	t		Yes	
Charger	p	Yes		Yes

Table 3. Data after manipulation

Table 4 shows some of our initial results:

Token	Contains DNT tag?	Contains translatable tag?	Comment
Black	b	c	Black and Decker brand vs. color
Case	b	n	Case Logic brand
Charger	md	ta	Dodge Charger car vs. device
RAM	md	ta	Dodge RAM pickup vs. RAM memory
Range	md	f	Range Rover brand vs. common word
Seat	ma	n	Car maker in Spain

Table 4. Results from NER

5.3. From Part of Speech data (POS)

This process is based on identifying when a word is used with different parts of speech in a certain content. It is very common for MT engines to make errors because of a word that is written in the same way, but can be a verb, a noun, or an adjective for example. While the English language does not have any difference for the usage of that word, other languages will have lots of variations for the different POS. Adjectives will have gender in Romance languages, and verbs will have a variety of forms. This brings again the concept that "translations will be different for the same word", and this may confuse the MT engine and affect the MT quality.

The premise for this process is that if a word is associated with two different POS types, there is a good chance that the word is polysemous (will have different translations).

We run a POS tagger on the content, and the result looks like this:

<S> Loring[Loring/NNP,B-NP-singular|E-NP-singular] was[be/VBD,B-VP] a[a/DT,B-NP-plural] dedicated[dedicated/JJ,dedicate/VBD,dedicate/VBN,I-NP-plural] artist[artist/NN,E-NP-plural] whose[whose/WP$,B-NP-plural] artistic[artistic/JJ,I-NP-plural] abilities[ability/NNS,I-NP-plural] and[and/CC,I-NP-plural] accomplishments[accomplishment/NNS,E-NP-plural] are[be/VBP,B-VP] beautifully[beautifully/RB,I-VP] shown[show/VBN,I-VP] in[in/IN,B-PP]

this[this/DT,B-NP-singular] book[book/NN,book/VB,book/VBP,E-NP-singular].[./.,</S>,O]

With some manipulation, we create a list with two columns: word and POS tag. We sort that list by word, and secondarily by tag, and then we are on our way to identify words that have more than one POS tag, as shown in Table 5:

Word	Tag	Comment
Accessory	J	
Accessory	J	
Accessory	N	accessory tagged as N (noun) and J (adjective)

Table 5. Data after POS tagging and manipulation

Different POS taggers will have different tags, but this process only requires:

- Creating a vertical list of words and tags (usually simple introduction of CR characters)
- Identifying a different part of the tag for nouns, adjectives and verbs (sometimes the first letter of the tag will be enough, as above)

We can also subtotal the list by words and tag, and we will then have information about the frequency that each word and tag occurs. This number indicates the candidates with better potential. One word may have a 60%/40% ratio between noun and verb, while another word may have a 99%/1% ratio. If the same proportion appears in the training data, the first situation will more likely confuse the MT engine than the second situation.

Table 6 below show some of our initial results:

Accessory	N	accessory tagged as N and J
Acted	V	acted tagged as V and J
Adapted	V	adapted tagged as V and J
Added	V	added tagged as V and J
Adhesive	N	adhesive tagged as N and J
Adjusted	V	adjusted tagged as V and J
Adore	V	adore tagged as V and N
Affected	V	affected tagged as V and J

Table 6. Results from POS

6. Quantification effect enhances relevance

The processes presented have a "quantification" effect on the meaning. A term could be polysemous and one of the meaning could be very rare. In practical terms, this would not be a significant polysemy case, because there is no volume for that meaning. The eBay data helps indicating how often a term has one meaning versus another, by connecting the meaning to a frequency number.

In queries, the frequency is defined by the category that follows the term. In NER, the frequency indicates how often each meaning appears in one category, but we can also look across categories. In POS, this effect also appears. In absolute terms, a certain word can be

tagged with several parts of speech. However, one of these POS may be very rare in the context being analyzed, so this variation would not appear in the results.

These are positive effects, because they introduce the frequency/relevance into the analysis and results, as opposed to an analysis based just on the absolute existence of multiple meanings or POS in a dictionary.

7. Conclusion

The processes described here are finding words with limited human effort, indicating that they are efficient. These words are valuable for eBay because they take into account the eBay context. For example, Fossil is a noun and a brand, but a dictionary would not contain the brand. So these processes are finding words in a way that could be difficult to find with other resources. There is also value in the "quantification" of how frequent these words are.

The methods for harvesting polysemous words presented here are only possible due to the wealth of linguistic data that eBay has. We hope that other companies that have data will find these ideas useful, and those who do not have data will feel inspired to create data and use it.