



“Durschnittskopfhörer” or the Translation of Sentiment

How far will machine translation take sentiment analysis technology?

Butler Hill Group, June 2010

Introduction

There has been a lot of buzz around sentiment analysis, a text analytics application that aims to extract and classify emotions, attitudes and opinions from unstructured text sources. Companies use sentiment analysis to categorize and make actionable what customers express about their brands, products, and services on social media sites, blogs etc. To date, sentiment analysis has been primarily focused on the US market, but global companies have global customers. To respond to the need for capturing “global sentiment”, some sentiment analysis providers are looking to machine translation as a cheap and fast way to extend their offerings into new languages.

We at Butler Hill wanted to explore how machine translation holds up in this context. Having worked with language technologies since our inception 16 years ago, we understand the strengths and weaknesses of machine translation. We designed an experiment to investigate the impact of machine translation on precision and recall of sentiment analysis. Using German and Japanese customer review data, we rated reviews first in their original language using in-market annotators, and then rated the machine-translated English versions using US-based annotators. Our findings reveal the challenges companies will face for both recall and precision of sentiment analysis technology when relying on machine translation.

This whitepaper provides valuable insights to anyone in the sentiment analysis industry with global customers or with a product roadmap to extend sentiment analysis to new markets.

About Sentiment Analysis

As the amount of user-generated content grows exponentially, developers of text analytics technologies are addressing the need to mine this unstructured data and turn it into actionable information. One exciting application of text analytics is sentiment analysis: Beyond understanding what people are talking about (the topic, specific entities, etc.), companies also want to know how they think and feel about that topic or those entities. A vice president for customer care might want to know what product shortcomings leave users most

frustrated; a chief marketing officer might seek customer feedback to a new branding campaign; and a product planner might want to pick up trends and attitudes amongst a target customer segment. Sentiment analysis refers to the automated extraction and classification of people’s opinions, attitudes and emotions using natural language processing (NLP) methods like entity extraction and shallow or deep parsing, which can be either statistical or rule-based. In addition to determining sentiments, the analysis may also extract the intensity of the sentiments expressed, their degree of polarity, as well as who the key opinion holders are.

Challenges for Sentiment Analysis

Sentiment analysis is an especially challenging application of text analytics, because opinions are more nuanced than facts. While some opinion sources include descriptive metadata that can provide context for analytical efforts (e.g. star ratings for product reviews), many sources have to be analyzed on the basis of the text itself.

The most salient clues for identifying sentiment are given by lexical choice (e.g. “His new movie stinks”, “Hated the color”, “Addictive taste”.) Lexical choice is not just domain-dependent but also context-specific on a topic level, in that an individual word can have different meanings depending on the specific product or service aspect being discussed: Compare “I didn’t sleep well because the hotel room was quite hot” *versus* “The design of the lounge is really hot”, which could appear in the very same user review of a particular hotel.

Beyond individual words, phrasings and the organization of the text contribute relevant information, as do modal operators (“If the bed were firmer, I would have slept well”) and punctuation (“You call this tasty??”).

In addition to general issues that come up when analyzing unstructured “noisy” text (misspellings, fractured grammar, abbreviations, slang, etc.), sentiments are often expressed through sarcasm, irony, highly individualized jargon, and a good dose of personality. Humans can easily differentiate between the literal meaning and the intended meaning in a sentence like “I REALLY dig a phone that drops calls all the time”, but imagine how hard this is for an automated system. Finally, consider semantic and pragmatic implications as in “I eventually managed to sleep comfortably” with the implication that it was actually quite difficult to do so.

As sentiment analysis technologies are deployed by global companies, the expansion of the technologies into non-US markets is a topic many are just beginning to confront. Some sentiment analysis developers claim that their technology is “language agnostic” and will successfully analyze non-English text that has been machine-translated into English: They believe that while machine translation is not perfect, the occasional missed sentiment or even incorrectly classified sentiment will not negatively impact the overall result set. Others

believe that in order to get a true representation of “native” feedback in the respective target language, one must adapt the sentiment analysis tools and incorporate knowledge about the lexical, morphological, syntactical and semantic structures of the language. There are also cultural differences in how sentiments are expressed across languages that affect the analysis process and outcome, though no one outside small academic circles has explored this aspect so far.

An Experiment

We at Butler Hill understand techniques for scaling language technologies like text analytics into global markets, and we have evaluated and helped improve machine translation technology for over a decade. We decided to take a closer look at the opportunity for, and effectiveness of, using machine translation in sentiment analysis by creating a human-annotated data set of sentiments and by exploring the degree of consistency between ratings for the original data and ratings for machine-translated data. We see this experiment as a first step toward developing best practices around truly globalizing sentiment analysis technology. We took a set of German and Japanese customer reviews from the respective Amazon sites, had them rated for sentiment by native speakers, then ran those sentences through Google Translator and had the machine-translated sentences rated by English annotators without German/Japanese knowledge. The ratings we asked annotators to apply were:

- positive
- negative
- positive/negative (for sentences that contain both positive and negative sentiments)
- neutral (for sentences that were meaningful but did not express a sentiment)
- not ratable (for sentences whose meaning could not be identified.)¹

¹We used 997 German and 588 Japanese sentences, extracted from Amazon customer reviews for BOSE head phones on the respective market's Amazon website. We used equal amounts of sentences from 1 and 2 star reviews (assumed to contain more negative sentiments overall), and 4 and 5 star reviews (assumed to contain more positive sentiments overall). The reviews were broken into individual sentences automatically, and randomized before the being handed off to annotators, meaning the annotators did not know what star rating a sentence came from and/or what preceded or followed an individual sentence.

Each German and Japanese sentence was rated by two native annotators, who then proceeded to resolve all initial rating differences through a manual arbitration step. We then ran all German and Japanese sentences through Google Translator and handed off the resulting sentences to English annotators without German or Japanese language skills; again, each sentence was rated twice, followed by manual arbitration. All annotators followed the same rating guidelines, which we prepared internally after conducting a trial annotation project on comparable data.

We wanted to find out whether and how the ratings would differ between the native and non-native speakers, whether the quality and quantity of such differences would vary across languages, and what the implications might be for an actual sentiment analysis application.

Findings

Machine translation impacts both quantity and quality of sentiment data from foreign sources

While we observed many interesting differences across the ratings for the German, Japanese and machine-translated English sentences, the most impactful ones concerned sentences that were rated as positive or negative (vs. neutral or not ratable) by native annotators. These are the most important sentences for a sentiment analysis application to extract and classify correctly:

- Missing positive or negative sentences and instead rating those neutral or not ratable means the system's recall is affected.
- Misclassifying a positive or negative sentence affects the system's precision.

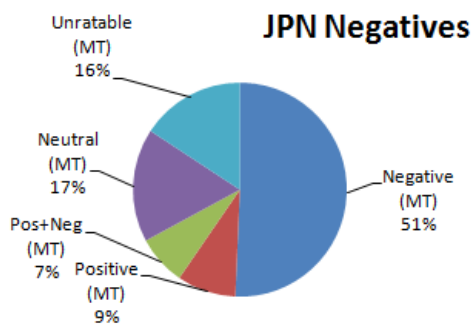
As an example for the latter, note the difference between the machine-translated version and the actual meaning of an original Japanese sentence:

- Japanese sentence: 基本的に高音はキレイです。 → Rated: positive
Human translation for reference: Basically the high note is beautiful.
- Machine translation: The treble is pretty basic. → Rated: negative

We found that in German, almost a quarter of the sentences that native annotators rated as positive or negative were rated differently by English annotators when presented with the machine-translated version: 23% of original German sentences rated positive were rated differently after being machine-translated, and 25% of original German sentences rated negative were rated differently after being machine-translated.

In Japanese, the numbers of discrepancies were even higher: 40% of original Japanese sentences rated positive were rated differently after being machine-translated, and 49% of original Japanese sentences rated negative were rated differently after being machine-translated.

Even more striking, for 9% of the sentences originally rated positive or negative by Japanese annotators, the ratings of the machine-translated versions were polar opposites: While the Japanese annotators rated the original sentence positive, the English annotator rated the machine-translated version negative, or vice versa.



Just over half of the sentences rated “negative” by Japanese annotators got a corresponding negative rating by English annotators. 9% of negative Japanese sentiments were rated “positive” by English annotators based on the machine translation output.

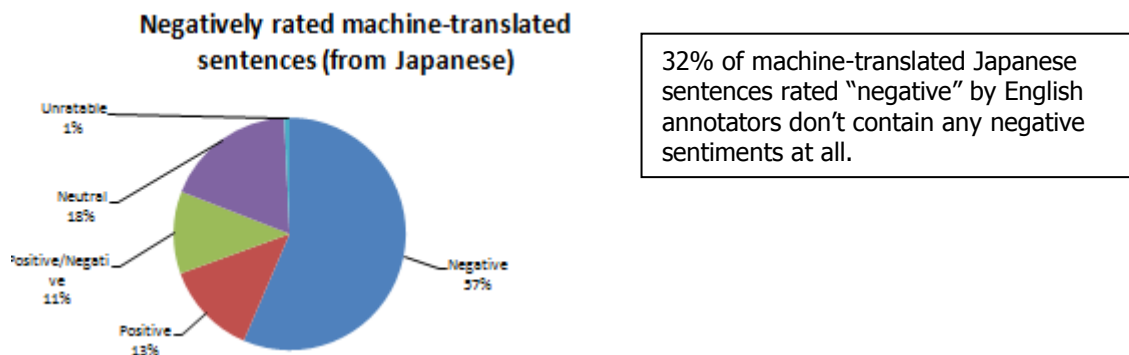
The impact to a sentiment analysis system is substantial: If these numbers are representative, *recall* for negative sentiments in Japanese is reduced almost in half as a result of utilizing machine-translated data rather than the original Japanese data. This represents a significant loss of input for clients who are usually most interested in the negative sentiments, as those are the main drivers for taking action in customer service, consumer feedback or PR scenarios.

Furthermore, if almost every tenth sentence considered negative gets misclassified as positive by English annotators of the machine-translated sentences, the system’s precision is likely to be impacted.

We also looked at the data from another angle, analyzing the machine-translated sentences that were rated “negative” by English annotators. Those are the sentences that a user of a text analytics system would pay the closest attention to: for example, the data that a vice president of customer care for an American company will want to review in order to understand the product shortcomings that leave Japanese users most frustrated. We wanted to see how relevant this set of sentences would be to anyone trying to create an actionable plan based on sentiment analytics.

We found that only 56% of machine-translated Japanese sentences and 74% of German sentences tagged as “negative” actually expressed purely negative sentiments in their original form. Accounting for sentences that were rated partially negative, our experiment points to a set of 32% (for Japanese) or 21% (for German)

sentences a user of a text analytics system would see in their high priority negative data set that are actually positive, neutral, or not ratable. This makes the data less conclusive and thus less actionable.



The goal of our experiment and analysis is not to critique the accuracy of machine translation systems. Machine translation does not aim –nor claim- to mirror the accuracy and nuances of human translation; rather, it is a terrific tool for getting the “gist” of content in a foreign language, and it has many highly valuable applications and uses for individuals and businesses. We do want to show the shortcomings of machine translation in a context where the data is “messy” to begin with, and where seemingly minor changes to the sentence structure, a different part of speech designation, or a choice of word sense, can lead not just to a different meaning but to a different category of meaning for sentiment analysis. Compare for instance the difference between “I stuck to this model”² (positive rating) vs. “I’m stuck with this model” (negative rating), caused by an incorrect switch from active to passive voice by the machine translation.

While the most critical issues introduced by machine translation are the misclassification of positive sentences as negative and vice versa, additional issues come up that affect recall and precision. Sometimes individual words or whole sentences remain untranslated and, unless the context makes it clear what sentiment is expressed, the sentence is lost to the system altogether. See for example “Otherwise, the Bose sound like a Durchschnittskopfhörer” (“Durchschnittskopfhörer” being a misspelled term for “average headphone”) or the short and not-so-sweet “FINGER WEG”, which translates to “stay away” and should have counted as a strong negative statement; neither sentence could be rated by English annotators.

² German original phrase: “Ich bin bei dem Modell geblieben”

The machine translation approach will have a different success rate across languages

We found many more rating differences across Japanese original vs. machine translated sentences than across the corresponding pairs for German. For Japanese, we also found that ratings for sentences in their native language were more *consistent across annotators* compared to the machine-translated sentences: Japanese annotators agreed on their ratings for 70% of the sentences before arbitration, while English annotators had same ratings for only 60% of the corresponding machine translated sentences. For German, the distribution of agreements vs. disagreements was comparable between native and machine-translated versions.

Further analysis is needed to determine whether this disparity is due to differences in the quality of the machine translation between Japanese and German, or whether it is more challenging to “translate” Japanese sentiment expression into English due to culture-specific ways of expressing sentiments that get lost or distorted through the machine translation process and the ratings by English annotators. One implication, however, is clear: A company extending sentiment analysis technologies into new markets has to look at each language individually. Even if machine translation meets the bar for one language, it might not work equally well for another language.

Discussion

As outlined above, using machine translation as a means of extending sentiment analysis to non-English text comes with a decrease in both recall and precision. Whether the decrease is critical depends on the specific application of the technology and a variety of surrounding factors. Some technologists argue that with large-scale data collection, the errors introduced may not be statistically significant: Statistically, the missed or misclassified sentences amongst billions of tweets and blog entries will just be noise in the system. If, however, the total amount of sentiment data available to a particular user of the applied technology (a business decision owner in a customer care team, PR agency, or marketing department) is small, errors in the sentiment classification will matter.

If the output of sentiment classification is actually shown to consumers, as for instance in online services that aggregate product or hotel reviews, the impact of showing a misclassified sentiment expression is much bigger than if it just surfaces to an internal analyst inside a company. One bad example may cause customers (or in-company executives) to distrust the application altogether.

Finally, if the data is sensitive or where a negative sentiment has implications beyond mere “customer dissatisfaction” (e.g., where personal data is involved, or sentiments expressed might have implications for public health or safety), machine translation might just be too risky in that too much gets “lost in translation”.

We believe that there will be contexts in which machine translation will help a business gain insights from non-English customer data, especially if an automated (machine translation assisted) classification phase is supplemented by manual confirmation, correction, and augmentation. For many contexts, however, true “native” sentiment analysis systems and processes might be needed, especially where clean data is scarce and where accurate classification of the sentiment drives important business decisions.

Global text analytics is too new a field for any established best practices. However, at this point we recommend that anyone considering an approach that translates non-English sentiment data into English for classification by a US English classifier, evaluates the errors introduced by machine translation, and determines if the error rate is statistically significant and acceptable for their intended application. Based on our findings, we are skeptical whether the machine translation approach yields the right data needed to make business decisions, and we did not even take cultural differences into consideration in our experiment.

Opportunities

As a team of linguists and language specialists, we are passionate about supporting language technology companies in their global endeavors. Text analytics in general, and sentiment analysis specifically, present countless interesting challenges, and we plan to explore these independently and in partnership with our clients. Our shared goal is to create and implement processes for extending technologies into new languages in a predictable and scalable way, and to assist with making the right tradeoffs between quality, speed, and cost. Talk to us to discuss how we may help you take your offerings into new markets!