Co-financed by the European Union
Connecting Europe Facility

A Case Study of Machine Translation in Financial Sentiment Analysis

Chong Zhang v-chong.zhang@lionbridge.com
Department of Linguistics, Stony Brook University

Matteo Capelletti Matteo.Capelletti@lionbridge.com
Lionbridge Technologies, Inc

Alexandros Poulis Alexandros.Poulis@lionbridge.com
Lionbridge Technologies, Inc

Thorben Stemann v-Thorben.Stemann@lionbridge.com
Lionbridge Technologies, Inc

Jane Nemcova Jane.Nemcova@lionbridge.com
Lionbridge Technologies, Inc

Abstract

The European research project Social Sentiment Indices powered by X-Scores (SSIX) intends to allow Small and Medium-sized Enterprises (SMEs) to take advantage of social media sentiment data for the finance domain. The project aims to overcome language barriers and realize a financial sentiment platform capable of scoring textual data in different languages.

Our approach to achieve this goal takes maximum advantage of human translation while keeping costs low by incorporating machine translation. In the long run, we intend to provide a tool that helps SMEs to expand into new markets by analyzing multilingual social contents.

In this paper, we investigate how sentiment is preserved after machine translation. We built a sentiment gold standard corpus in English annotated by native financial experts, and then we translated the gold standard corpus into a target corpus (German) using one human translator and three machine translation engines (Microsoft, Google, and Google Neural Network) which are integrated in Geofluent to allow pre-/post-processing. We then conducted two experiments. One meant to evaluate the overall translation quality using the BLEU algorithm. The other intended to investigate which machine translation engines produce translations that preserve sentiment best.

Results suggest that sentiment transfer can be successful through machine translation if using Google and Google Neural Network in Geofluent. This is a crucial step towards achieving a multilingual sentiment platform in the domain of finance. Next, we plan to integrate language-specific processing rules to further enhance the performance of machine translation.

1. Background

Over the past two years, Lionbridge has been involved as a leading industrial partner in the European funded SSIX project (Social Sentiment Index, 2015 - 2018). During the project (which will be completed in February 2018), we have developed a platform for detecting opinions about stocks, companies and their products as expressed in social media and other media sources. For example, we can extract content from Twitter, StockTwits, news, company blogs, etc and analyze sentiment associated to each content.

In Lionbridge, we conceive the SSIX platform as a supporting tool for our sales representatives. Our goal is to make it easier to detect the following aspects:

- What are the needs of our customers
- What prospects may be entering within our areas of expertise
- What are the weak and strong points of our competitors

We consider such knowledge as strategic to trigger appropriate action in real time. For example, we can track customers' needs on social media and adjust our services accordingly in real time; we can detect events that are relevant to our interests and deal with them strategically.

In the past, a sales representative would need to search different sources in an *accessible locale* to find relevant discussion of new products or market updates. This was done in the past manually to a large extent. Such manual approach may not be ideal for many reasons: it is prone to missing information, slow in response time, and expensive in terms of human labor.

Now the SSIX platform offers the possibility to partially automate the search. It allows search terms and media channels to be defined, and it notifies users of changes amongst public opinion. It allows us to see what people say about products and companies in real time. Futhermore, this is not restricted to a specific language and locale. Thanks to the integrated technology of Lionbridge GeoFluent (GeoFluent, Lionbridge Inc.), we can overcome the language barrier and provide financial sentiment analysis across languages.

2. Introduction

One of the primary targets of the SSIX project is sentiment analysis in the financial domain across multiple languages. The work has started with English, where a three-way validated sentiment gold standard has been developed and has been used to train the sentiment classifier. The work on English can rely on several available resources, such as text normalization tools, polarity lexica and distributed word representations that allow the development of a sentiment classifier for English to be based on pre-existing resources.

The work started with building a three-way validated sentiment gold standard corpus for English (Hürlimann et Al., 2016). Three experts in the domain of finance annotated the English corpus manually, and their sentiment scores were reconciled for consistency. This gold standard corpus was used to train and test the SSIX sentiment classifier.

Addressing languages different from English, however, is a more complex issue that raises a series of questions. Resources for other languages may neither be as readily available, nor as good in quality. This raises the question whether it is possible/sufficient to rely on the resources we have for English to address sentiment classification for other languages. Suppose, as it is in fact the case, that we want to develop a sentiment classifier for German when we already have a working version for English. Is there a way to capitalize on the resources developed for English to create a classifier for German?

To answer this question, we suggest at least three approaches:

1. Create a gold standard corpus for German from the ground up, manually annotate and cross review it, and then train the new classifier on it. We call this the *Native* approach.

2. Take the English sentiment gold standard corpus, translate it (either manually or automatically) to German, and train the German classifier on it. We call this the *Derived* approach.

3. Use machine translation to convert the German input to English, and feed the English translations to the English classifier. We call this the *Direct Translation* approach.

The three approaches obviously differ in quality, efficiency and costs. Each approach has its advantages and disadvantages, which are briefly outlined below.

2.1. The Native Approach

Building a new Gold Standard corpus from scratch, as in the Native approach, is expensive, but potentially very rewarding. The most prominent benefit is that no translation is taking place and the native expert judgments are on “first hand” data. Creating such a gold standard is both costly and time-consuming, as we need more than one annotator (at least 3) to agree on the sentiment of each piece of text in order to ensure good quality data. Considering that the sample should contain several thousands of tweets and that a domain like Finance needs judgments made by specialists, the cost may quickly skyrocket. On the other hand, the only variable in the Native Approach is the agreement of the annotators, provided their individual domain knowledge and familiarity with the exchange media (tweets) does not lead to vastly different sentiment scores for the same data. *Due to the conditions of its design and implementation, we could assume that once available, such a gold standard would be the standard against which any other approach should be benchmarked.*

2.2. The Derived Approach

In this approach, instead of building a new corpus and annotating it manually, we use the already existing English language gold standard and translate it to German. This approach presupposes that a statement with positive sentiment in English remains positive in German and vice-versa for negative judgments. Several translation methods are available: It can either be done manually, via machine translation, or in a hybrid way, using computer aided translation tools or post-translation review by human translators. We can also take advantage of the fact that only some words are sentiment-bearing thus targeting these words in context for optimal translation and ignoring the rest.

If we use human translation, the task of creating a translated GS will be cheaper than the creation of a native GS, in the sense that one domain expert will probably be enough, where previously three were needed. Certainly, the cost and time decrease drastically when using machine translation, but the resulting data, especially in a technical domain such as finance, may be of lower quality. Machine translation could, for instance, systematically map an English term to a German term which is synonymous in some other domain, but which is not relevant to the financial domain.

A human-reviewed machine translation is surely the safest approach if one wants to speed up the process and keep costs limited. This may actually reveal error patterns in the translation that can be fixed in post-processing.

2.3. The Direct Translation Approach

Instead of training a new classifier on German data, we translate the German input text to English and feed it to the English classifier. Clearly, translation here can mean only machine translation, as we will be dealing with large amounts of input data to be processed in real time. This approach can also add further costs as machine translation on large amounts of data comes at a cost.

The translation-based approaches in 2 and 3 face a number of issues related to the domain and the specificity of the text involved. Spelling errors, uncommon abbreviations and rhetorical text are all extra challenges that need to be tackled.

Input normalization and output optimization are strategies that can be pursued to improve the quality and accuracy of the translation. First, we may remove elements like repeated characters or delete unknown strings. During post-analysis of translated material, we can map common MT mistakes to the desired output, for instance, terms that need a specific translation in the domain of reference. There is a large range of operations that can be performed – some language-specific, some more general. In this respect, **GeoFluent** [2] is specifically designed not only to support automatic translation but also in preparing the input and correcting the output of the translation process (pre- and post-processing of the data).

3. Setup

The work discussed in this paper is a contribution to the Derived and Direct Translation approaches.

Within the scope of the SSIX project, we built a sentiment gold standard corpus for English, annotated by native experts from the domain of finance (Hürlimann et Al., 2016). The gold standard corpus was translated into a target corpus in German by a domain expert. At the same time, it was also translated into German by three machine translation engines. These are Microsoft, Google, and Google Neural Network, which are integrated in Lionbridge GeoFluent [2]. We used GeoFluent to introduce pre-/post-editing, such as DO-NOT-TRANSLATE rules to tackle special financial terms and text normalization rules.

In SSIX, we intend to take maximum advantage of human translation while keeping the cost low by incorporating the machine translation component. Our objective is to use manually translated data as a benchmark and examine machine translation outputs: their quality and preservation of sentiment in the financial domain.

A crucial prerequisite for our approach is that the sentiment of the gold standard corpus can be transferred to the target corpus after translation. If the sentiment is lost after translation, either by human or by machine, we cannot use our previous research results, i.e. the English sentiment classifier, and implement either the Derived approach or the Direct Translation approach. The only viable option left would be the Native approach, which is bound to have high costs. As a result, to meet the prerequisite and make decisions for further actions, we must investigate the impact of machine translation on the sentiment quality of the gold standard corpus. We have conducted two experiments to study how machine translation influences sentiment, as discussed below.

4. Experiment 1

The first experiment was designed to find out the quality of each machine translation engine. In this experiment, we selected a sample of 700 English tweets from Twitter and StockTwits relative to the financial domain. This data set was selected for its clarity in expressing sentiment. For example, textual data that did not offer valuable information such as containing only URLs was filtered out to reduce noise.

During the experiment, this sample was translated into German simultaneously by one human translator and the three machine translation engines mentioned above, namely Microsoft, Google, and Google Neural Network, as integrated in Lionbridge GeoFluent. The human translator is a native speaker of German and a domain expert in finance.

To evaluate translation quality for the three machine translation engines, we calculated their BLEU scores (Koehn et al., 2007; for source code see References). Using human translation as the reference, the three machine translations were each compared to the human translation to see how close they are to the professional human translation[1].

The results are summarized in the table below. They suggest that Google and Google Neural Network performed better than Microsoft on 1-gram, and Microsoft performed better than Google and Google Neural Network on 2-grams, 3-grams, and 4-grams.

Engine	1-gram	2-grams	3-grams	4-grams
Microsoft	0.901470798	0.865873923	0.786125067	0.684824095
Google	0.963509145	0.846959705	0.728174371	0.605465403
Google Neural Network	0.963340387	0.846025029	0.727096883	0.604167208

Table 1. BLEU score for machine translations

The 1-gram is used to assess how much information is retained after translation. Clearly Microsoft has lost more information than both Google and Google Neural Network. Among 2-grams, 3-grams, and 4-grams calculations, 4-grams is believed to be the most correlated with judgements made by native speakers of the target languages (Papineni, K., et al., 2002).

[1] We understand that BLEU score is meant to evaluate translations on a corpus level. However, due to time and resource limitations, at this stage we can only investigate the current data sample size. We consider expanding our data size and reduplicating this experiment in order to confirm our results in future.

Our results suggest that Microsoft produced the most similar translations to human translator. Google and Google Neural Network performed more poorly in comparison.

However, we must notice that the BLEU algorithm was not sufficient for our purposes because it only evaluates translation quality in the respect of approximating human translation. Since the purpose of SSIX is to build a sentiment platform, *we consider the quality of translation is the best when there is minimal discrepancy in sentiment between the original texts and the translations*. Using our criterion, we need to explore the sentiment preservation. That is why we conducted Experiment 2.

5. Experiment 2

4.1 Experiment Design

For Experiment 2, we selected a subset of the previous sample (N = 200). We had to reduce the size of our sample because Experiment 2 required much more human resources than Experiment 1. To keep the time and expense cost under control, we chose a subset of the prevous sample.

This experiment was designed to investigate whether translations (regardless of whether they came from human translators or machine engines) can maintain the sentiment from the original texts. As the first step, we recruited two German financial domain experts and they assigned sentiment to all four translations. The experts were kept away from the original English texts and their sentiment.

The sentiment scores assigned by the domain experts ranged from 1 to 10, 1 being the most negative, and 10 being the most positive. If the assigned pair of scores for a certain line of text diverged from each other for more than 2 points (including 2), we asked a third domain expert to evaluate the text again and chose the more appropriate sentiment score from the two alternatives.

For example, the human translator translated a certain tweet into German: *"Der miterlebte Fortschritt ist echt atemberaubend." - Stifel Analyst, nachdem er Teslas Fabrik zum vierten Mal gesehen hat $TSLA https://t.co/nD7KECoM6V*

Its original English tweet is: *The progress witnessed is truly stunning." - Stifel analyst after seeing Tesla's factory for the fourth time $TSLA https://t.co/nD7KECoM6V*

One of our domain experts assigned the German translation a sentiment score of 3, and the other assigned it a 10. Since there was a big gap between the two scores, the third domain expert evaluated the translation, and chose 10 from the pair of 3 and 10. As a result, the sentiment score for this tweet is 10.

4.2 Results and Discussions

After the data were evaluated and reconciled in the above way, we performed some statistical analysis on the results. We used a mixed linear regression model, which was implemented with the lmer4.0 package in R (Federico et al., 2014; Guzman et al., 2012). Compared with a linear regression model, a mixed effects model can explicitly model invidual characteristics. In our design, we used the item as a random intercept to capture the variance of each translated item to maximize the differences we could find between compared sets.

We are mainly concerned with the following two questions:

- Do human translations preserve sentiment?
- Does machine translation preserve sentiment?

To answer the first quesion, we need to compare the sentiment of the English gold standard corpus with the sentiment of human translation. If there was no significant difference between the sentiment scores of English gold standard and human translation, we would know the sentiment did not change too much; if a significant difference was found, then the sentiment is already lost in human translations.

After calculating our data set, results showed that there was no significant difference between the sentiment of English gold standard and human translation (Figure 1). In other words, the difference between gold standard sentiment (mean = 5.67 4) and human translation sentiment (mean = 5.536) was not large enough for us to draw the conclusion that they are different on a statistical level. This proves that human translation can preserve sentiment from the original texts. The results are what we desire to see because human translation is believed to be more reliable than machine translation. If human translation could not preserve sentiment, it is unlikely that machine transltion can.

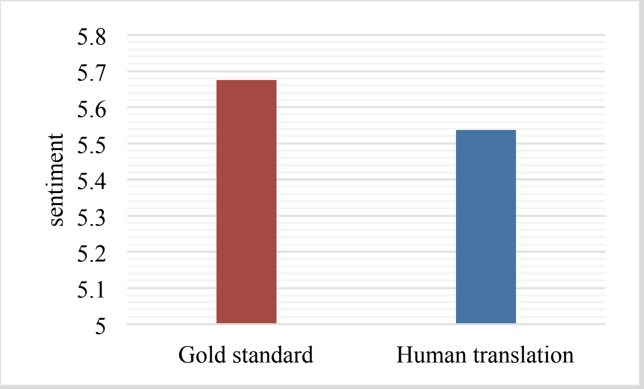

Figure 1. Sentiment Comparison: Gold standard vs. Human

Next, we try to answer the second question and assess the performance of machine translation engines on sentiment preservation. We compared the sentiment of the English gold standard with the sentiment of machine translations. Our results suggested that there were significant differences between the three pairs, i.e. English gold standard vs. Microsoft, English gold standard vs. Google, and English gold standard vs Google Neural Network (Table 2).

Engine	t-value	p-value
Microsoft	t = -3.574	p < .001
Google	t = 2.038	p < .05
Google Neural Network	t = 3.101	p < .01

Table 2. Results for Sentiment Comparison (Gold standard vs. Machine)

The visualization of the result can be found in Figure 2[2]. Here Microsoft shows stronger diversion from the original sentiment in the gold standard, and Google produced the sentiment that was the closest to the original.

We also notice that compared to the gold standard sentiment mean, both human and machine translations have sentiment with lower means. At least two factors attribute to this fact. One is that translations have "neutralized" sentiment, drawing its mean closer to the grand mean (i.e. 5.5) because translations always lose information to an extent. The other is due to our domain experts. We used different groups of domain experts for annotating sentiment of English and German data, who are English and German native speakers respectively. Our German annotator could be more conservative or negative in assigning sentiment scores.

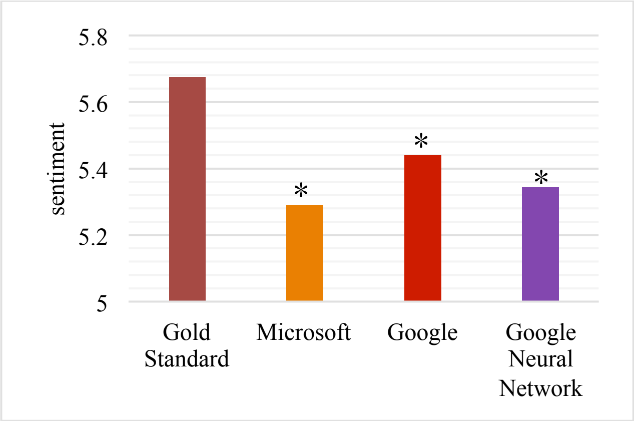

Figure 2. Sentiment Comparison: Gold standard vs. Machine

These results indicate that translations generated by machine engines are not of the desired high quality and look to be at risk of losing or distorting sentiment. However, they do not imply that machine translation is without merit. Since we have established that human translation is successful in preserving sentiment, we can use human translation as the benchmark to compare machine translations. If the sentiment assigned to a given machine translation engine does not deviate significantly from that of human translation, we can conclude that the engine has produced sentiment scores comparable to human translation.

The three comparisons discussed above showed that there are significant differences between the sentiment of human translation and Microsoft, which indicates that the Microsoft engine did not produce translations whose sentiment was alike to human translation (Table 3). The visualization is provided in Figure 3.

[2] The * on top of the bars indicated significance

Engine	t-value	p-value
Microsoft	t = -2.16	p < .05

Table 3. Results for Sentiment Comparison (Human vs. Machine)

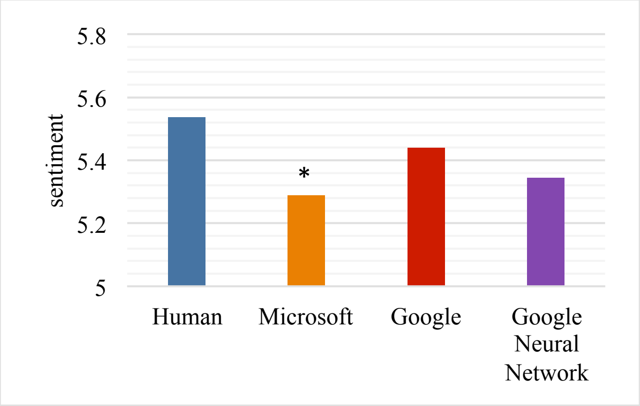

Figure 3. Sentiment Comparison: Human vs. Machine

Crucially, there was no significant difference between the sentiment scores of human translations and both Google and Google Neural Network. This means that the sentiment scores from Google and Google Neural Network does not differ significantly from human translation. This proves that these two engines' performance was in line with human performance, and consequently in these cases, sentiment can be considered as successfully preserved.

6. Conclusion

In this paper, we provide evidence that sentiment can be preserved after translation of an English gold standard corpus into German by machine engines, namely Google and Google Neural Network when they are integrated in GeoFluent. With this prerequisite fulfilled, we can either use the Derived approach to convert English data to another language and subsequently train a sentiment classifier on that data. Alternatively, we can use the Direct Translation approach to transfer multilingual data to English and use our already built English sentiment classifier. As these approaches do not need a human translator, time and costs can be greatly reduced, without an apparent, major loss in quality for the purposes of sentiment analysis. This is a crucial step for building an affordable multilingual sentiment platform in the domain of finance, to overcome the language barriers and help SME to analyze multilingual social content.

We have many directions for further research in the future that go from the integration of more language-specific processing rules in GeoFluent to enhancing the performance of machine translation, to benchmarking financial sentiment classifiers trained with Native and Derived approaches.

ACKNOWLEDGMENTS

This work is funded by the SSIX Horizon 2020 project (Grant agreement No 645425).[3]

References

Federico, M., Negri, M., Bentivogli, L., Turchi, M., & Kessler, F. F. B. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. In *EMNLP* (pp. 1643-1653).

GeoFluent (Lionbridge Inc.) http://www.lionbridge.com/GeoFluent/

Guzman, F., & Vogel, S. (2012). Understanding the Performance of Statistical MT Systems: A Linear Regression Framework. *Proceedings of COLING 2012*, 1029-1044.

Hürlimann M., Davis B., Cortis K., Freitas A., Handschuh S., Fernández S. (2016 September). *A Twitter Sentiment Gold Standard for the Brexit Referendum*. Paper presented at the Proceedings of the 12th International Conference on Semantic Systems, Leipzig, Germany.

Koehn, P., & Schroeder, J. (2007, June). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the second workshop on statistical machine translation* (pp. 224-227). Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.

Social Sentiment Index, 2015 – 2018, https://ssix-project.eu/

Source code for calculating the BLEU score:
http://www.nltk.org/_modules/nltk/translate/bleu_score.html

[3] The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.