Does 'well-being' translate on Twitter?

Laura K. Smith[1] **Salvatore Giorgi**[1] **Rishi Solanki**[2] **Johannes C. Eichstaedt**[1]
H. Andrew Schwartz[3] **Muhammad Abdul-Mageed**[4] **Anneke Buffone**[1] and **Lyle H. Ungar**[5]
[1]Department of Psychology, University of Pennsylvania
[2]Electrical and Systems Engineering, University of Pennsylvania
[3]Computer Science, Stony Brook University
[4]Library, Archival and Information Studies, University of British Columbia
[5]Computer and Information Science, University of Pennsylvania
lasm@sas.upenn.edu, sgiorgi@sas.upenn.edu

Abstract

We investigate whether psychological well-being translates across English and Spanish Twitter, by building and comparing source language and automatically translated weighted lexica in English and Spanish. We find that the source language models perform substantially better than the machine translated versions. Moreover, manually correcting translation errors does not improve model performance, suggesting that meaningful cultural information is being lost in translation. Further work is needed to clarify when automatic translation of well-being lexica is effective and how it can be improved for cross-cultural analysis.

1 Introduction

Interest in sentiment analysis spans academic and commercial domains, with wide-ranging applications (Pang and Lee, 2008; Liu, 2012). While the majority of tools for sentiment analysis have been developed for English text, ideally sentiment and emotion could be analyzed across many languages. Does one need to build models for each language of interest, or can models be applied cross-culturally? More generally, how do cultures differ in the language they use to express sentiment and feeling?

Sentiment in resource-poor languages has commonly been assessed by first translating text into English and then applying an English sentiment model (Mohammad et al., 2016). This approach is economical and efficient, as building each model of interest in every target language is resource-intensive. Yet it is not clear how much culturally specific information and accuracy are lost in the translation process, and specifically how this varies across languages, cultures, linguistic content, and corpora (e.g., social media vs. news). While extensive work has demonstrated that automatic machine translation (MT) methods are competitive when translating opinion in news and blogs, less research has examined the translation of sentiment on social media, and specifically on Twitter, known for its restriction of individual exchanges to short samples of text (140 characters) and informal language. Moreover, research has not focused on translating subjective well-being specifically.

Beyond sentiment, this paper investigates how expressions of personal well-being translate between English and Spanish on Twitter. We have English and Spanish speakers annotate Tweets in their native language for five components of subjective well-being (positive emotion, engagement, positive relationships, meaning, and accomplishment) (Seligman, 2011). We then compare how well models trained and tested in the same language compare to (a) models developed in one language, and then translated (using Google Translate) to the other language (e.g., how well English models translated to Spanish work on Spanish Tweets) and (b) how well models developed in one language work on Tweets translated from another language (e.g., how well English models work on Tweets translated from Spanish to English).

2 Related Work

There is a vast literature on sentiment analysis which space precludes us from surveying; see (Liu, 2012)

for an excellent overview. A small but rapidly growing camp is developing methods to estimate personality and emotion, asking "how does she feel?" rather than "how much does she like the product?" (Mohammad and Kiritchenko, 2015; Park et al., 2014). In social media, the well-being of individuals as well as communities has been studied, on various platforms such as Facebook and Twitter (Bollen et al., 2011; Schwartz et al., 2013; Eichstaedt et al., 2015; Schwartz et al., 2016).

2.1 Translating sentiment

Past work has, on the whole, regarded state-of-the-art automatic translation for sentiment analysis optimistically. In assessing statistical MT, (Balahur and Turchi, 2012) found that modern SMT systems can produce reliable training data for languages other than English. Comparative evaluations between English and Romanian (Mihalcea et al., 2007) and English and both Spanish and Romanian (Banea et al., 2008) based on the English MPQA sentiment data suggest that, in spite of word ambiguity in either the source or target language, automatic translation is a viable alternative to the construction of models in target languages. Wan (2008) shows that it is useful to improve a system in a target language (Chinese) by applying ensemble methods exploiting sentiment-specific data and lexica from the target language and a source language (English). More recent work has examined how sentiment changes with translation between English and Arabic, also finding that automatic translation of English texts yields competitive results (Abdul-Mageed and Diab, 2014; Mohammad et al., 2016). However, translated texts tend to lose sentiment information such that the translated data is more neutral than the source language (Salameh et al., 2015).

It is less obvious how well expressions of emotion or subjective well-being translate between languages and cultures; the words for liking a phone or TV may be more similar across cultures than the ones for finding life and relationships satisfying, or work meaningful and engaging.

2.2 Well-being

In contrast to classic sentiment analysis, well-being is not restricted to positive and negative emotion. In 2011, the psychologist Martin Seligman proposed PERMA (Seligman, 2011), a five-dimensional model of well-being where 'P' stands for positive emotion, 'E' is engagement, 'R' is positive relationships, 'M' is meaning, and 'A' is a sense of accomplishment. PERMA is of interest to this translation context because while the 'P' dimension maps relatively cleanly onto traditional conceptions of sentiment (i.e., positive and negative emotion), PERMA also includes social and cognitive components which may be expressed with more variation across languages and cultures. In recent work, Schwartz et al. (2016) developed an English PERMA model using Facebook data. In this paper, we adopt a similar method when building our message-level models over Tweets.

Governments around the world are increasingly dedicating resources to the measurement of well-being to complement traditional economic indicators such as gross domestic product. Being able to measure well-being across regions is not only becoming more important for institutions and policymakers, but also for private sector entities that want to assess and promote the well-being of their organizations and customers. This raises the importance of translation, given that resources for the measurement of well-being are disproportionately available in English.

3 Methods

We collected Spanish data using the Twitter API, gathering 15.3 million geolocated Tweets between September and November 2015 using a latitude/longitude bounding box around Spain. This set was reduced to messages containing only Spanish using the Language Identification (LangID) Python package (Lui and Baldwin, 2012). We restricted to messages with an 80% or higher Spanish confidence score as given by LangID. This resulted in 6.1 million Tweets from 290,000 users. We selected 5,100 random messages from this set for annotation. English Tweets were similarly collected using the Twitter API, restricted to the US, and filtered to be (primarily) in English.

3.1 Annotating message-level data

Amazon's Mechanical Turk (MTurk) was used to annotate the 5,000 random English (American)

Tweets[1]. CrowdFlower, an online crowdsourcing platform similar to MTurk, but more widely used in Europe, was used to annotate our 5,100 random Spanish Tweets[1]. As the Tweets exclusively came from Spain, raters were restricted to fluent Spanish speakers who live in Spain.

On both MTurk and CrowdFlower, separate annotation tasks were set up for each of the 10 PERMA components (positive and negative dimensions for the 5 components). Workers were given the definition of the PERMA construct, directions on how to perform the task, and were presented with an example annotation task. During the task workers were asked to indicate "to what extent does this message express" the construct in question on a scale from 1 ("Not at all") to 7 ("Extremely"). Directions were presented in English for the English task, and in Spanish for the Spanish task. The Spanish instructions were translated manually from English by a bilingual English-Spanish speaker and verified by an additional bilingual speaker.

In the English task, two raters assessed each message. If the raters disagreed by more than 3 points, a rating was obtained from a third rater. It proved more difficult to get raters for the Spanish task, even on CrowdFlower. In some cases we were unable to obtain even a single annotation for a given Tweet and PERMA component.

3.2 Developing weighted lexica

Tweets were tokenzied using an emoticon-aware tokenizer, 'happy fun tokenizer'[1]. We then extracted unigrams and bigrams from each corpus, yielding vocabularies of 5,430 and 4,697 'words' in English and Spanish, respectively. The presence/absence of these unigrams and bigrams in each Tweet were used as features in Lasso (L1 penalized regression) (Tibshirani, 1996) models to predict the average annotation score for each of the crowdsourced PERMA labels. Separate models, each consisting of regression weights for each term in the lexicon, were built for each of the ten (five positive and five negative) PERMA components in both English and Spanish[1]. Each model was validated using 10-fold cross validation, with Pearson correlations averaged over the 10 positive/negative PERMA components. Results are presented in Table 1. The models were then transformed into a predictive lexicon using the methods described in (Sap et al., 2014), where the weights in the lexicon were derived from the above Lasso regression model.

Model	***r***
Spanish	0.36
English	0.36

Table 1: Performance as measured by Pearson *r* correlation averaged over the 10 positive/negative PERMA components using 10-fold cross validation.

3.3 Translating the models

We used Google Translate to translate both the original English and Spanish Tweets and the words in the models. We also created versions of the translated models in which we manually corrected apparent translation errors for 25 terms with the largest regression coefficients for each of the 10 PERMA components (the top 250 terms for each model).

3.4 Comparative evaluation

We evaluated how well the different models worked, computing the Pearson correlations between message-level PERMA scores predicted from the different models and the ground-truth annotations. Lexica were built on 80% of the messages and then evaluated on the remaining 20%. Figure 1 shows test accuracies. Comparing the English and Spanish source language and machine translated models, we observe substantially better performance when models were built over the same language they are applied to, i.e., using models built in Spanish to predict on Spanish Tweets. Translating the models (e.g., translating an English model to Spanish and using it on Spanish Tweets) or translating the Tweets (e.g., translating Spanish Tweets to English and using an English model) work substantially less well, with translating the Tweets giving marginally better performance than translating the models. Finally, we translate both the model and Tweets, giving slightly better performance than translating the Tweets alone. Complete PERMA lexica were then built over the entire message sets for public release.

3.5 Error Analysis

To quantify the errors in translation, we took the 25 top-weighted words in each component of the

[1] Available at www.wwbp.org.

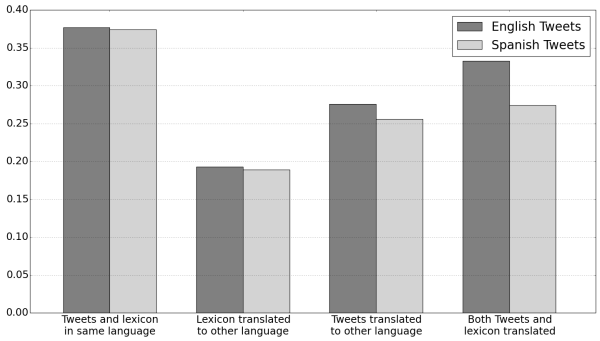

Figure 1: Performance (Pearson *r* correlation) between ground-truth annotations and predicted lexica scores averaged over the 10 PERMA components.

source lang	correct trans	missing terms	opp sign	weight diff
English	83%	81%	0.5%	6.9%
Spanish	74%	91%	0.0%	4.8%

Table 2: Summary of translation errors. Percentages are averaged over the 10 PERMA components. *Source lang* is the language of the model which was translated, *correct trans* is the percentage of correct automatically translated words, *missing terms* is the percentage of correct automatic translations within the 250 top terms that did not appear in the top 250 words of other source model, *opp sign* is the percentage of terms whose sign switched between models, and *weight diff* is the percentage of terms whose weights between the two models were off by a factor of two.

PERMA	term	weight (en)	weight (es)	% chg
POS_M (en)	mundo* (world)	0.42	-0.18	143
NEG_A (en)	odio** (hate)	0.29	2.19	87
NEG_M (en)	nadie*** (no one)	0.23	0.24	4.2
NEG_R (es)	sad** (triste)	1.70	0.0012	100
NEG_P (es)	hate*** (odio)	1.81	1.75	3.3

Table 3: Examples of specific errors. Error types are denoted by asterisks: * denotes a change in sign, ** denotes the largest change in weight and *** denotes the smallest change in weight per source model. Language listed under each PERMA category is the language of the source model that was translated. The *% chg* column is percentage change relative to the larger weight. For clarity, under each term we include its translation.

PERMA lexicon (250 terms total) and manually translated them with the help of a native Spanish speaker. The manual translations were then compared against the automatic translations. Out of the top 25 words we calculated the percentage of correct automatic translations (when manual and automatic translations matched) and averaged the percentages across positive and negative PERMA components. The average percentage of correct translations is listed in Table 2 as *correct trans*.

These correctly translated terms were then compared to the terms in the opposite source model (i.e., after translating English PERMA to Spanish, we compared the translations with Spanish PERMA). We calculated the percentage of the top 250 translated words missing in the 250 top words of the source lexicon for each PERMA component and averaged over the 10 components. This value is reported in Table 2 as *missing terms*. For terms that appeared in both the translated and source lexica we compared their respective weights, calculating both percentage of terms in which the weights were of different signs and percentage of terms with substantially different weights. Again, these percentages were averaged over the 10 PERMA components. Percentages are reported in Table 2 as *opp sign* and *weight diff*, respectively. To be considered "substantially different" the two weights must differ by a factor of 2. It is worth noting that at no point were the translated and source weights equal (within a tolerance of 10^{-5}).

We then looked at the errors term by term. Out of the 500 terms considered (top 250 words per source language) only one term weight changed signs between models: "mundo" (world). The weight for this term in the translated English to Spanish model was 0.42 whereas the weight in the Spanish model was -0.18, amounting to a 140% change. Next, for each source model we report terms with the largest and smallest differences in weight. These terms and weights are reported in Table 3. The language abbreviation ("en" or "es") listed under each PERMA component is used to denote the source language we translated from. For example, (en) indicates that we started with English PERMA, translated it into Spanish and then compared to Spanish PERMA.

4 Discussion

The difference in performance between source and machine translated models can be attributed to a few main problems. First, the translation might be inaccurate (e.g., from our corpus, "te" is not in fact "tea"). We manually corrected translation errors in the prediction models with the help of a native Spanish speaker, but found that translation error accounts for marginal discrepancy between the source language and machine translated models.

A second source of errors are translations which are technically accurate, yet do not translate culturally. For instance, even though "andaluces" translated correctly into "Andalusians," "Andalusia" (an autonomous community in Spain) does not invoke the same cultural meaning in English as it does for Spaniards. A machine would be hard-pressed to translate "Andalusia" into a relevant region within the U.S. that might invoke similar popular sentiment. Although Spanish and American people share some holidays, musicians, and sports heroes, many of these differ (e.g., "Iker Casillas" is not well known in the U.S. and "La selectividad" may be similar to the "SATs," but this is not captured in MT).

A third source of error stems from cultural differences, with certain topics resonating differently cross-culturally. For instance, when comparing the highest weighted positive terms across PERMA, religious language (e.g., "god," "blessed") appears in English but not Spanish, fitting with the popular notion that Americans are more religious than Europeans. Spanish PERMA's positive emotion component contains multiple highly weighted instances of laughter; none have high weights in the English model. Highly weighted English negative emotion terms are marked by general aggression (e.g., "kill," "stupid") whereas the highest weighted Spanish terms include derogatory terms for disliked people (e.g., "douchebag," "fool"). The American positive relationship component is marked by words like "friend" and "friends," while "sister" is weighted more highly in Spanish PERMA.

Note that this is fundamentally a problem of domain adaptation rather than MT, as our error analysis revealed that the majority of top-weighted terms were exclusive to one source model. Different cultures use different words (or at least vastly different word frequencies) when revealing the same kind of well-being. Exploring where the sentiment around a similar concept diverges across languages can provide insight to researchers studying cross-cultural variation.

4.1 Limitations

This work has significant limitations. First, the English and Spanish annotation processes, though kept as similar as possible, were not identical; annotations were gathered on different platforms, and due to our difficulty in recruiting Spanish raters, our total annotations per message varied across tasks. Additionally, the models were built over relatively small corpora of 5,000 English Tweets and 5,100 Spanish Tweets. These Tweets came from different time periods, which may further reduce similarity between the Spanish and English corpora. Finally, our method does not account for the presence of various sub-cultures within the United States and Spain.

5 Conclusion

In this work, we investigated how well expressions of subjective well-being translate across English and Spanish Twitter, finding that the source language models performed substantially better than the machine translated versions. Moreover, manually correcting translation errors in the top 250 terms of the lexica did not improve model performance, suggesting that meaningful cultural information was lost in translation.

Our findings suggest that further work is needed to understand when automatic translation of language-based models will lead to competitive sentiment translation on social media and how such translations can be improved. Cultural differences seem more important than language differences, at least for the tasks we studied here. We expect that language indicators of personality and emotion will similarly translate poorly, but that remains to be studied.

Acknowledgments

The authors acknowledge support from the Templeton Religion Trust (grant TRT-0048) and Bioibérica.

References

Muhammad Abdul-Mageed and Mona T Diab. 2014. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, LREC, pages 1162–1169.

Alexandra Balahur and Marco Turchi. 2012. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA, pages 52–60.

Carmen Banea, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 127–135.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, ICWSM, pages 450–453.

Johannes C Eichstaedt, H Andrew Schwartz, Margaret L Kern, Gregory Park, Darwin R Labarthe, Raina M Merchant, et al. 2015. Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2):159–169.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, ACL, pages 25–30.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 976–983.

Saif M Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326.

Saif M Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Greg Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, et al. 2014. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108:934–952.

Mohammad Salameh, Saif M Mohammad, and Svetlana Kiritchenko. 2015. Sentiment after translation: A case-study on arabic social media posts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, pages 767–777.

Maarten Sap, Greg Park, Johannes C Eichstaedt, Margaret L Kern, David J Stillwell, Michal Kosinski, et al. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing*, EMNLP, pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, et al. 2013. Characterizing geographic variation in well-being using tweets. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM.

H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, et al. 2016. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 516–527.

Martin EP Seligman. 2011. *Flourish*. Free Press, New York, NY.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 553–561.