Machine Translation as an Alternative to Language-Specific Dictionaries for LIWC

Yuying Ye
Independent researcher
y.ye.yuying@gmail.com
Twitter: @YuyingYe

Peter Boot
Huygens ING
peter.boot@huygens.knaw.nl

Linguistic Inquiry and Word Count (LIWC) (Tausczik & Pennebaker, 2010) is a text analysis program that calculates the percentage of words in a given text that fall into specific categories (social, sentiment, cognitive, biological, and more). The categories are defined in an English-language dictionary. In order to apply the program on non-English text, its dictionary has been translated into multiple languages (e.g. Boot et al. 2017; Meier et al., 2019).

We propose an alternative approach for applying LIWC to non-English text: using machine translation (MT) to translate the text into English and then apply the English LIWC dictionary. We evaluate whether this approach produces better or poorer results than analysing text with a translated dictionary. We see this as a contribution to conference theme 2: evaluating a practice of quantitative analysis using LIWC.

In several contexts, MT has been shown to be useful for extending the scope of English-language research tools. It has been used to create NLP-resources and tools for low-resource languages, for example, in subjectivity analysis (Banea et al. 2008) and sentiment analysis (Balahur & Turchi 2012). Also, MT translated text can produce similar or better results in sentiment analysis (Araujo et al., 2016) and topic modelling (De Vries et al., 2018). Specifically for LIWC, manual and machine translation give comparable results (Windsor et al., 2019), which didn't consider, however, the use of a translated dictionary.

In our experiment, we use two open-source MT systems, Joshua 6[1] (Post et al., 2015) and Fairseq[2] (Ott et al., 2019), and Google Translate. We use pre-trained MT models to facilitate replicability. We include three language pairs, Joshua 6 on Dutch-English, German-English and Spanish-English; Fairseq on German-English and Google Translate on Dutch-English, which also facilitates comparison among different MT architectures. The method is shown in Figure 1.

We evaluate the alternative method with TED Talk subtitle parallel data on Opus (Tiedemann, 2016), because it is representative speech data and covers an extensive range of topics. The parallel corpus was translated by human volunteers (Cettolo et al., 2013). We use the MT models to translate the non-

[1] https://cwiki.apache.org/confluence/display/JOSHUA/Language+Packs

[2] https://github.com/pytorch/fairseq/blob/master/examples/translation/README.md

English texts into English. Next, the (translated) texts are analysed by available versions of LIWC dictionaries: for Dutch-English, the 2007 human-translated LIWC dictionary and 2015 machine-translated dictionary; for German-English, 2001 and 2015 LIWC dictionaries; for Spanish-English, 2007 LIWC dictionary. We compare the LIWC output using correlation and effect sizes for each of the LIWC categories (see Figure 2 for an example of the evaluation). We also create parallel displays of the texts with highlighted LIWC hits to investigate what causes the differences.

Provisionally, MT seems to lead to better results in the three language pairs than language-specific dictionaries. While we do encounter errors in the translated dictionaries, the main cause of discrepancy between the two procedures seems to be the fact that many homonymies are specific to languages. English 'belief' is in the religion category, but the equivalent 'geloof' in Dutch is not, because it is also used as a form of the verb 'geloven' ('believe'), mostly in a non-religious context. Translating the sentence to English resolves the ambiguity. That may also lead to wrong categorisations: the word 'soul' in English is classified as religious, and so references to the music genre will also count in that category. In Dutch, the music genre is also called 'soul,' but is not in the religious category. Translating the text to English leads to wrongly counting the word as religious. Similarity to the English results is thus not the same as correctness. We will publish a deeper analysis of the different results of the two procedures, which may depend on the language, the corpus and the quality of the dictionaries.

In this talk we will discuss the technical details on utilising the MT models and show preliminary evaluation results. We also go into the technical process on applying MT and the practicalities that DH researchers might encounter.

References

Araujo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016, April). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (pp. 1140-1145).

Balahur, A., & Turchi, M. (2012, July). Multilingual sentiment analysis using machine translation?. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis* (pp. 52-60). Association for Computational Linguistics.

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 127-135).

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics, 6(1)*, 65–76. https://doi.org/10.1075/dujal.6.1.04boo

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., & Federico, M. (2013). Report on the 10th IWSLT evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*.

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417-430.

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). "LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015. PsyArXiv(a).

Ott, M., Edunov, Sergey., Baevski, A., Fan, A., Gross, S., Ng, Nathan., Grangier, D., & Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *Proceedings of NAACL-HLT 2019: Demonstrations*.

Post, M., Cao, Y., & Kumar, G. (2015). Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical Linguistics*

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29(1)*, 24–54. https://doi.org/10.1177/0261927X09351676

Tiedemann, J. (2016). OPUS--Parallel Corpora for Everyone. *Baltic Journal of Modern Computing*, 384.indsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PloS one, 14*(11).

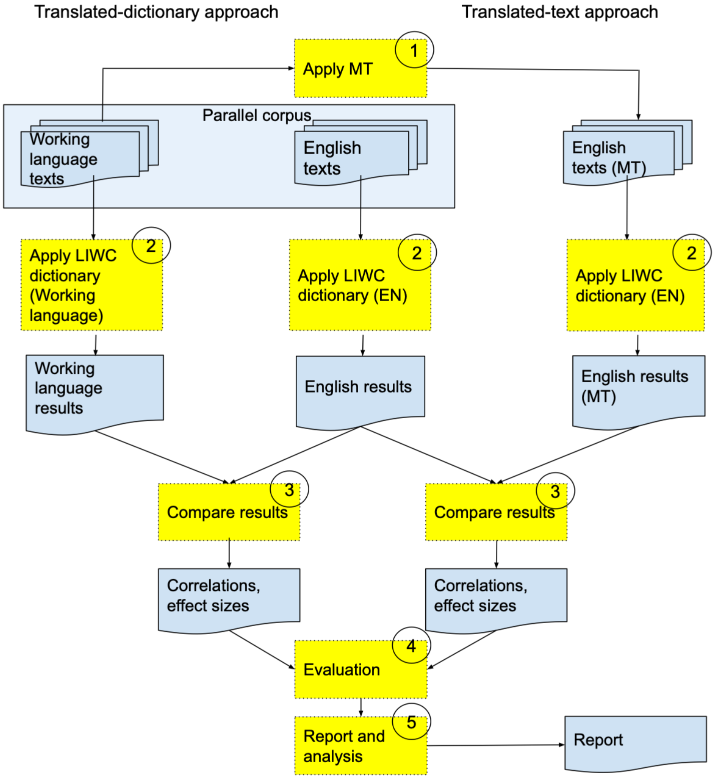

Figure **1** Experiment process.

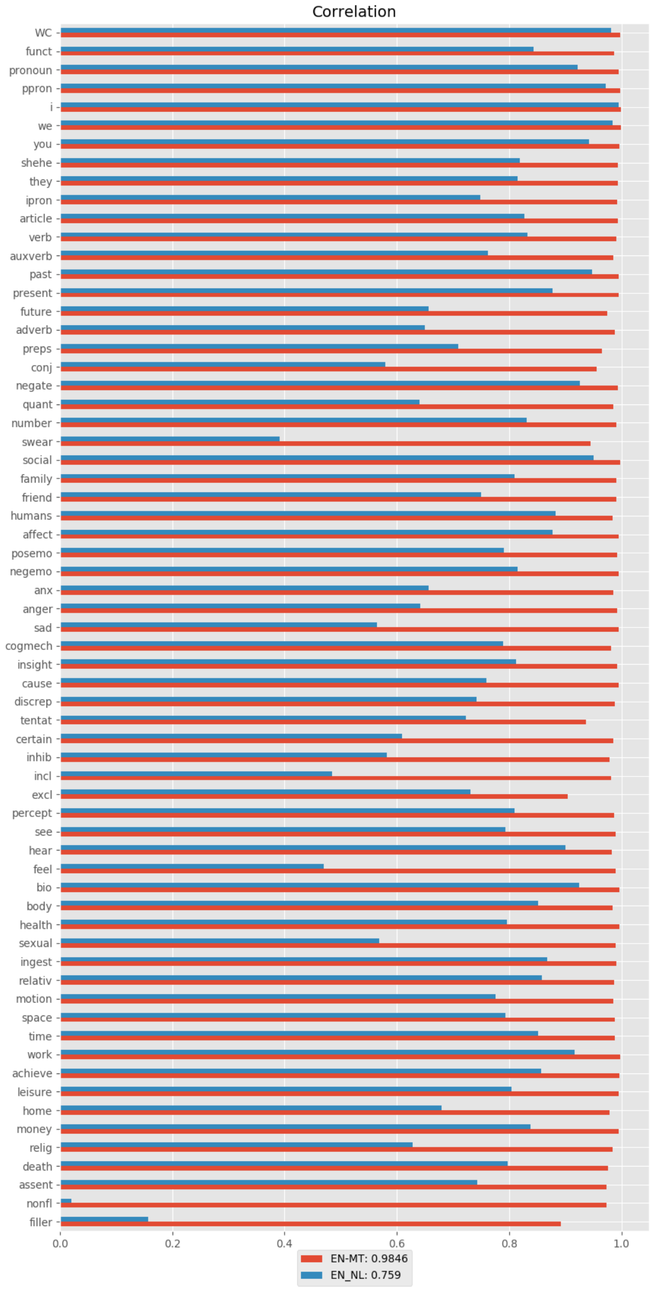

Figure 2 Correlations for the different categories. In red correlations between English gold standard and MT English, in blue English gold standard and Dutch. The average correlations is shown at the bottom.