VEGA: A Virtual Environment For Exploring Gender Bias vs. Accuracy Trade-Offs In AI Translation Services

Mariana Bernagozzi[1], **Biplav Srivastava**[2], **Francesca Rossi**[1], **Sheema Usmani**[1]
[1]IBM, NY, USA
[2]University of South Carolina, Columbia, USA
{Mariana.Bernagozzi, Francesca.Rossi2, sheema.usmani}@ibm.com, biplav.s@sc.edu

Abstract

Machine translation services are a very popular class of Artificial Intelligence (AI) services nowadays but public's trust in these services is not guaranteed since they have been shown to have issues like bias. In this work, we focus on the behavior of machine translators with respect to gender bias as well as their accuracy. We have created the first-of-its-kind virtual environment, called VEGA, where the user can interactively explore translations services and compare their trust ratings using different visuals.

Introduction

Machine translation services are widely available AI services. But do people see bias in their outputs? Are they willing to tolerate accuracy loss as a trade-off for absence of bias? Many studies have shown issues with AI services including those involving natural language processing (NLP) and machine learning techniques like machine translators (Blodgett et al. 2020).

Bias in computational systems (devices, application programming interfaces) is an impediment for adoption and is of increasing importance as apps become cognitive and interact with people. Racial, sexual, and religious biases, for example, can appear. Public trust in AI services is not guaranteed. This is true for transactional stateless services, such as machine translators, as well as for interactive stateful services, such as conversation agents (e.g., chatbots). Bias, hate speech, information leaking, lack of accuracy, etc. interfere with trust. In previous work, we proposed methods to rate primitive AI services and their sequential composition (Srivastava and Rossi 2020, 2018). The methods specially focused on machine translators and gender bias. Based on this, in the current work, we demonstrate VEGA, a tool to explore and visualize the trust ratings of machine translation services, providing comparative views of gender bias and accuracy.

Background

There has been previous work to assess bias in translators. In (Prates, Avelar, and Lamb 2018), the authors test Google Translate on sentences like "He/She is an Engineer" where occupation is from U.S. Bureau of Labor Statistics (BLS). They compare frequency of female, male and gender-neutral pronouns in the translated output and compares with BLS data about expected frequency. In another paper (Font and Costa-jussà 2019), the authors look at a transformer architecture for machine translation put in the Open Neural Machine Translation (Open NMT) translator[1] and two debiasing word embeddings. They consider sentences of the form: "I've known her/him/ $\prec$ proper noun $\succ$ for a long time, my friend works as {a, an} $\prec$occupation$\succ$." They consider English to Spanish and look at the form of friend used based on occupation. They make a list of 1019 occupations available[2].

In previous work, we have made progress on rating text-based AI services and their sequential composition:

1. Rating for translators (Srivastava and Rossi 2020, 2018), with a focus on gender bias, and
2. Rating for conversation agents (Srivastava et al. 2020) using combined rating of multiple trust issues

System Overview

For VEGA, three machine translators have been considered. We will refer to them as T1, T2, and T3; even if they are translators publicly available, since the focus of this study is the rating method and its visualization, rather than a comparison of the behavior of three specific translators. For each translator, we considered the outputs between English and 4 other languages: Arabic, Spanish, French, and Portuguese.

The system consists of a series interactive panels that are presented to the user during the different steps that comprise the flow of the system. Also, the system features a progress bar at the top to indicate previous, current and next steps and arrows to navigate between the steps. Below, we describe the details of each panel.

Panel 1: Translation

The purpose of this panel is to allow the user to explore the text outputs of the different translation services for a set of predefined sentences. Sentences can be translated from English to 4 different languages mentioned above and vice

[1]http://opennmt.net/; accessed 18 Sep 2020.

[2]At: https://github.com/joelescudefont/genbiasmt; accessed 18 Sep 2020.

versa. The direction of the translation can be inverted by clicking on the double arrows button.

The texts to be translated are made up of two sentences containing one gender place-holder each. The phrases follow the format: *≺Gender≻ is a ≺Occupation-Performer≻. ≺Gender≻ is a ≺Occupation-Performer≻*. We chose this two-sentence format because we wanted a text that could include both genders. This allows us to expose in a more articulate way the possible bias translation issues. For the gender, we use either He or She. For the occupation, we use a list of occupations from a public site [3]. An example is *She is a Florist. He is a Gardener.*

When translating from English, we present the user a list of 19 texts, in English, that follow the format described above. The translations produced by the three translators are shown at the bottom of the panel.

When translating back to English from any of the other 4 languages, the user is presented with the 19 triplets we obtained from translating the original 19 sentences, and their corresponding translations back to English are shown at the bottom of the panel now. A screenshot of this panel is presented in Figure 1.

Panel 2: Gender Bias Binary Rating

We tested the three translators for gender bias and Panel 2 shows our assessment. For each (language, translator) pair, the assessment is binary, with a tick denoting absence of gender bias in the translator service and a red cross denoting presence of gender bias in the translator service. The overall gender bias rating for a translator service is computed by taking the worst bias rating over the four languages.

Panel 3: Detailed Bias Classification

On this panel, we present a more fine-grained visualization of our gender bias rating, showcasing the three levels of bias ratings proposed in one of our previous works (Srivastava and Rossi 2018). The three ratings are:

Unbiased (U): This is the best rating and it means that the system not only does not introduce bias, but it does not even follow the bias of the input data, and is instead able to compensate for possible bias in the input data.

Data-Sensitive Biased (DSB): The system does not introduce bias but it follows whatever bias is present in the input data.

Intrinsically Biased (IB): The system introduces bias even when the input data is unbiased.

The overall rating is computed by taking the worst bias rating over the four languages. The language can be changed by clicking on the tiles to see how the translators perform for the different languages. The overall rating is shown at all times, using a horizontal dashed line over the bars.

Panel 4: Trade-Offs (Bias Rating vs. Accuracy)

On this panel, we show both the gender bias and the accuracy of the translators. Using this visualization we can reason on the possible trade-off between these two important dimensions of the quality of a translator. Again, the language can be changed by clicking on the tiles and the overall rating is shown using a horizontal dashed line over the bars. In the same fashion, overall accuracy rating is computed by taking the worst accuracy rating over the four languages. A screenshot of this panel is presented in Figure 2.

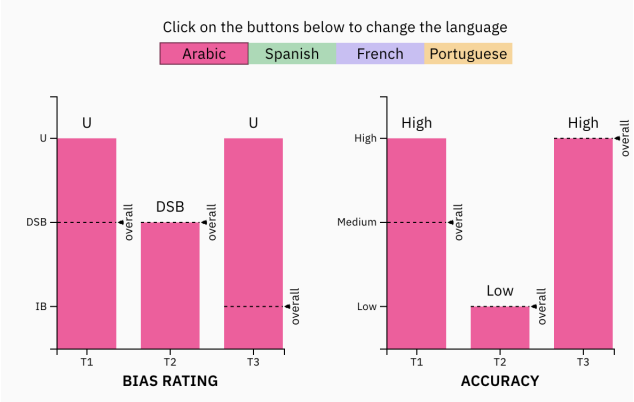

Figure 2: Gender bias vs. accuracy panel.

[3] http://www.vocabulary.cl/Basic/Professions.htm

Figure 1: Interactive translation panel: English to French.

Conclusion

To our knowledge, VEGA is the first tool to explore and compare gender bias and accuracy for machine translation services. We have built the system as a web application that allows users to interact with visuals to understand the behavior of the translators and provide feedback on their preferences. The system can be extended easily to incorporate more translators, languages, sentence forms and trust dimensions like bias on the basis of religion and race (via support of the rating system).

VEGA is available from this link: http://vega-live.mybluemix.net

References

Blodgett, S. L.; Barocas, S.; III, H. D.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey

of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Font, J. E.; and Costa-jussà, M. R. 2019. Equalizing Gender Biases in Neural Machine Translation with Word Embeddings Techniques. *CoRR* abs/1901.03116. URL http://arxiv.org/abs/1901.03116.

Prates, M. O. R.; Avelar, P. H. C.; and Lamb, L. C. 2018. Assessing Gender Bias in Machine Translation - A Case Study with Google Translate. *CoRR* abs/1809.02208. URL http://arxiv.org/abs/1809.02208.

Srivastava, B.; and Rossi, F. 2018. Towards Composable Bias Rating of AI Systems. In *2018 AI Ethics and Society Conference (AIES 2018), New Orleans, Louisiana, USA, Feb 2-3*.

Srivastava, B.; and Rossi, F. 2020. Rating AI Systems for Bias to Promote Trustable Applications. In *IBM Journal of Research and Development*.

Srivastava, B.; Rossi, F.; Usmani, S.; and Bernagozzi, M. 2020. Personalized Chatbot Trustworthiness Ratings. In *IEEE Transactions on IEEE Transactions on Technology and Society. Earlier version at https://arxiv.org/abs/2005.10067*.