Analysis of Shortcut Learning Features in Natural Language Processing
(Technical Paper)

Machine Translation Technology: The Advantages and Limitations of Machine Translators in the Academic Community
(STS Paper)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Wan Li

November 1, 2021

Technical Team Members:
Hanjie Chen
Andrew Wang

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Yangfeng Ji, Department of Computer Science

Bryn E. Seabrook, Department of Engineering and Society

Introduction

Natural language processing (NLP) is a branch of artificial intelligence centered around how computers learn to comprehend language through text and oral communication (*What Is Natural Language Processing?*, 2021). NLP originated from Machine Translation (MT), which is the field of computer science that attempts to translate content from one language to another. MT technologies such as Google Translate have greatly evolved over the last 15 years, yet they are still widely inaccurate and often cannot translate an input's true meaning. The inaccuracies become more problematic as society becomes more dependent on technologies that claim to understand human language.

Oftentimes, NLP models do not actually learn the language they are trying to understand and instead rely on data statistics and correlation between inputs and labels. When the wrong correlations are used, NLP models may utilize shortcut features that use the wrong logic to solve the targeted task. For example, a model can rely on how often a word is used to determine whether it has a positive or negative connotation, or it can only analyze the last sentence of a paragraph instead of taking into account the whole input (Du et al., 2021). These models can produce high accuracy results, but the model itself would use the wrong features for predictions, resulting in flawed logic connecting inputs to outputs. Thus, these models would perform poorly on more challenging test conditions (Geirhos et al., 2020). The final technical deliverable focuses on identifying shortcut features in NLP models and creating a method to mitigate them so the model can use the correct logic to solve the problem at hand.

Technologies based on understanding language are widely used in the modern world despite their inaccuracies. Applications include translating educational resources (Kordoni et al., 2016), helping international students learn the host country's native language, and assisting students who study foreign languages. However, these benefits bring problems of academic integrity and the need for language development and diversity (Groves & Mundt, 2021). The final STS deliverable will focus on how MT can become a driving force in society, specifically in education, as well as the limitations and the societal expectations surrounding the technology.

Technical Topic: Shortcut Mitigation in NLP Models

Many problems related to difficult machine learning problems are symptoms of shortcut learning. Shortcut learning, as Geirhos et al. puts it, are "decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions, such as real-world scenarios" (2020). Essentially, shortcuts reveal a mismatch between the model's intended solution and the learned solution.

There are a few studies that have looked at specific shortcuts and how they can be mitigated. Du et al. found that NLP models have a strong preference for features located at the head of the distribution, or features that are introduced first or most frequently, resulting in a lack of attention to important information that may be at the tail of the distribution (Du et al., 2021). The researchers proposed a mitigation framework that regularizes the distribution and suppresses the model from making overconfident predictions when the shortcut is deemed highly apparent.

A case study centered around visual commonsense reasoning found that question answering models, or models that try to determine the right answer choice to a question, do not perform as intended because they often cheat by using the frequency of words in the answers, or look at how many words the question and each answer choice have in common (Ye & Kovashka, 2021). The researchers approached the problem with an iterative masking technique, which ran the model several times while deleting, or masking, a different part of the input each time to determine which words are important and which words or phrases the model needs to ignore while making predictions.

Although there is ongoing research about specific shortcuts found in NLP like high-frequency word association (Geirhos et al., 2020) and negation bias (Mendelson & Belinkov, 2021), it is impossible to remove shortcut features entirely from machine learning models. Researchers can only mitigate known shortcut features but unknown ones may never be discovered. The technical final deliverable will search for unknown shortcuts (Du et al., 2021). First, a list of known possible shortcut features will be listed and defined. The model will be trained first to see which known shortcuts are present, and then the solutions to those specific shortcuts found in previous references will be implemented to mitigate the found shortcuts. After all applicable known shortcuts are accounted for and mitigated, the model might still not perform as intended. If so, then there must be at least one additional shortcut feature that was not identified from the start. The model will be analyzed to identify what the additional shortcut features are, and then the deliverable will focus on mitigating the newly identified features. A goal of the technical project includes identifying shortcut features that are not superficial. In other words, the project wants to determine if there are shortcut features related to the semantics or meaning of the input words, as opposed to unrelated factors like word frequency.

The accuracy of the model in terms of learning the intended solution can be measured using a model's interpretation. Specifically, the concept of integrated gradients will be used to analyze how a model determines an output based on the input information (Sundararajan et al., 2017). Interpretations shed light onto what logic a model uses to solve a problem. If that logic seems to match how humans generally approach the problem, then the model has done well learning the intended solution. Without interpretations, the model is a black box: the user can only see the inputs and outputs of the model, but there is no information about how they connect. With the help of integrated gradients to determine the model's interpretation, the technical project researchers can determine the correctness of the model and identify which shortcuts may still be present. When the interpretation matches the intended solution, then the model has "learned" correctly and can solve the problem with the correct logic. The final deliverable will be a technical report that describes the background, related works, methods, results, discussion, and conclusion for mitigating unknown shortcuts in NLP using interpretations.

STS Topic: The Effect of MT technology on Education

Google Translate is the most popular Machine Translation (MT) technology in the modern world and currently supports translation for 109 different languages (*Compare Translate Features for Each Language - Google Translate*, n.d.). According to a survey conducted by Google in 2017, the MT product is 85% accurate and is still lacking in many aspects ("How Accurate Is Google Translate | Updated Review 2021," 2021). Google Translate is typically better at translating European languages than most others (Aiken, 2019) but even then, the technology errors most while trying to understand the actual meaning behind the words being translated (Ghasemi & Hashemian, 2016).

Despite being error-prone, Google Translate's accuracy is similar to the minimum standard necessary for university admission at many institutions (Groves & Mundt, 2015). The accuracy brings some ethical implications of using MT in higher education, since MT can cause problems of academic integrity (Yang & Wang, 2019) and change the need for language development (Groves & Mundt, 2021). On the other hand, MT is used to help international students learn the native language through self-studying (Bahri & Mahadi, 2016) and plays a large role in translating educational resources (Kordoni et al., 2016) or online educational content (van Rensburg et al., 2012) into other languages for global educational use. The stakeholders of MT in education include students studying foreign languages, international students, foreign students, professors, and the wider academic community. Although there are some cons to incorporating MT into education, the potential benefit this technology holds is an incredibly large win for education worldwide.

Some STS scholars may argue that MT could soon become a technologically deterministic technology. The concept of technological determinism originated from Thornstein Veblen and argues that society is shaped and driven by technology (Smith, n.d.). This concept is used often as an antonym for social constructivism, which argues that technology is driven by societal expectations. Once MT has evolved to be more accurate and pick up on the semantics of words, it will likely drive the academic community to create rules surrounding the usage of MT, as well as the economic impact of mass translating academic material. The potential benefits of MT will change how students and professors approach courses and assignments, which is why some may argue that MT will become a technologically deterministic technology. However, one of the shortcomings of the concept of technological determinism is that it does not have a timeframe; it simply states that technology drives society, from the start until it is replaced by a better technology.

Technological momentum, a concept made popular by Hughes, introduces a time aspect to the duality of technological determinism and social constructivism (Marx & Smith, 1994). Hughes argues that developed societies tend to lean towards technological determinism, whereas less-advanced societies tend to adopt social constructivism. Technological momentum is good for putting MT in education into perspective: in less advanced societies, society drives MT because it is typically smaller, so social norms revolving around MT matter more than the technology itself. In advanced societies, MT are more technologically deterministic because they are able to drive society and can easily be accepted into society's daily routines.

Researching how MT technology affects education is important because MT is rapidly evolving to become more accurate. If discussions are not held about how to incorporate MT into the academic community, problems of academic integrity will continue and students will not be able to effectively utilize MT in their education. It is also important to explore the limitations of MT in the context of its necessity. If MT is so accessible and accurate, is it necessary for students to learn foreign languages at all (Gally, n.d.)?

Methodologies

Research question: How is Machine Translation technology affecting education?

To answer this question, the methodology of historical case study will be used. Information will be collected and analyzed to understand how MT has affected specific institutions that require students to learn languages, and how MT has affected international students studying abroad. Sources will also be used to understand the ethical concerns regarding the incorporation of MT into the educational community, as well as its current limitations and what the technology still needs before being widely used in the academic world. Finally, research will be done on how MT is impacting large societies to change the structure of their educational system, whether that is removing the requirement to learn foreign languages or creating policies surrounding the implementation of MT technology.

The sources gathered will mostly be published accounts of how MT has affected different academic communities. The keywords used will be variations of "machine translation", "education", "higher education", "translation technology policies", "machine translation limitations", and "societies and machine translation". These keywords are important because they look for the connection between MT and education, as well as the ethical concerns and limitations of the technology. This will create a good library of information to explore the current societal effects of MT and how it can affect society in the future after overcoming some limitations.

Conclusion

The technical deliverable is a final report that will focus on what shortcut features were identified that were previously unknown, as well as what algorithms were implemented to mitigate those newly discovered shortcut features. The significance of the shortcuts will be analyzed to determine if they are superficial or pertain to the machine's understanding of the languages' semantics. The report will also include which known shortcut issues were identified at the start and mitigated using solutions from related works. Once completed, there will be insights into what shortcut features exist for NLP and their mitigation strategies. This will allow NLP models to improve the reasoning behind their answers, and also ensures that they are truly learning the material as intended. As a result, the models will have a higher accuracy and perform better because the model is working with the intended logic.

The STS deliverable is centered around how the rise of machine translation technology affects education worldwide. The final paper will analyze the ethical concerns regarding the use of MT for learning purposes, as well as the current limitations of the technology and its predicted advantages. Once the research is completed, the academic community will be able to better understand the impact MT technologies have on education. The research will facilitate discussions regarding the societal expectations surrounding the use of this technology, as well as what limitations need to be imposed. Nations will also be able to better understand the importance of languages within their academic systems and what they should expect from MT technologies moving forward, since MT will only continue to improve.

References

Aiken, M. (2019). An Updated Evaluation of Google Translate Accuracy. *Studies in Linguistics and Literature*, *3*, p253. https://doi.org/10.22158/sll.v3n3p253

Bahri, H., & Mahadi, T. S. T. (2016). Google Translate as a Supplementary Tool for Learning Malay: A Case Study at Universiti Sains Malaysia. *Advances in Language and Literary Studies*, *7*(3), 161–167.

Compare Translate Features for Each Language—Google Translate. (n.d.). Retrieved October 3, 2021, from https://translate.google.com/intl/en/about/languages/

Du, M., Manjunatha, V., Jain, R., Deshpande, R., Dernoncourt, F., Gu, J., Sun, T., & Hu, X. (2021). Towards Interpreting and Mitigating Shortcut Learning Behavior of NLU models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 915–929. https://doi.org/10.18653/v1/2021.naacl-main.71

Gally, T. (n.d.). *Machine Translation and English Education in Japan*. 13.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, *2*(11), 665–673. https://doi.org/10.1038/s42256-020-00257-z

Ghasemi, H., & Hashemian, M. (2016). A Comparative Study of “Google Translate” Translations: An Error Analysis of English-to-Persian and Persian-to-English Translations. *English Language Teaching*, *9*(3), 13–17.

Groves, M., & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, *37*, 112–121. https://doi.org/10.1016/j.esp.2014.09.001

Groves, M., & Mundt, K. (2021). A ghostwriter in the machine? Attitudes of academic staff towards machine translation use in internationalised Higher Education. *Journal of English for Academic Purposes*, *50*, 100957. https://doi.org/10.1016/j.jeap.2021.100957

How Accurate is Google Translate | Updated Review 2021. (2021, March 5). *The Language Doctors*. https://thelanguagedoctors.org/how-accurate-is-google-translate/

Kordoni, V., van den Bosch, A., Kermanidis, K. L., Sosoni, V., Cholakov, K., Hendrickx, I., Huck, M., & Way, A. (2016). Enhancing Access to Online Education: Quality Machine Translation of MOOC Content. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 16–22. https://aclanthology.org/L16-1003

Marx, L., & Smith, M. R. (1994). *Does Technology Drive History?: The Dilemma of Technological Determinism*. MIT Press.

Mendelson, M., & Belinkov, Y. (2021). Debiasing Methods in Natural Language Understanding Make Bias More Accessible. *ArXiv:2109.04095 [Cs]*. http://arxiv.org/abs/2109.04095

Smith, M. R. (n.d.). Technological Determinism in American Culture. . . *The MIT Press.*, 35.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. *ArXiv:1703.01365 [Cs]*. http://arxiv.org/abs/1703.01365

van Rensburg, A., Snyman, C., & Lotz, S. (2012). Applying Google Translate in a higher education environment: Translation products assessed. *Southern African Linguistics and Applied Language Studies*, *30*(4), 511–524. https://doi.org/10.2989/16073614.2012.750824

What is Natural Language Processing? (2021, August 16). https://www.ibm.com/cloud/learn/natural-language-processing

Yang, Y., & Wang, X. (2019). Modeling the intention to use machine translation for student translators: An extension of Technology Acceptance Model. *Computers & Education*, *133*, 116–126. https://doi.org/10.1016/j.compedu.2019.01.015

Ye, K., & Kovashka, A. (2021). A Case Study of the Shortcut Effects in Visual Commonsense Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3181–3189.