Automatic Tagging of Stack Overflow Questions Using Global Vectors and Deep Learning

Mentors:

Alina Lazar (Computer Science and Information Systems) alazar@ysu.edu

Bonita Sharif (Computer Science and Information Systems) bsharif@ysu.edu

Students:

Hannah Senediak (Computer Science and Information Systems) hesenediak@student.ysu.edu

Alexandra Ballow* (Mathematics and Physics) alballow@student.ysu.edu

Youngstown State University

Department of Computer Science and Information Systems

*Department of Mathematics and Department of Physics and Astronomy

Motivation

The Stack Overflow website, one of the most popular "question and answer" sites, is an essential and a growing resource among the community of coders all over the world. Software developers use it to post questions and answers related to programming and computer science problems they need to solve. Questions such as seeking input on some efficient and time-saving methods of coding a particular program, getting help on solving various bottlenecks in coding are commonly seen. Over the years the website has slowly evolved into a large free repository of knowledge. Currently, the site receives around 8000 questions per day, and includes 16 million questions, 24 million answers and 66 million comments all available to download in a data dump collection. The data is made publically under the Creative Commons cc-by-sa 3.0 license. Given the availability and size of the dataset many researchers from fields such as information retrieval, text mining and machine learning have been working with this textual dataset to solve different problems [1].

Tags have been used to identify online postings not only by Stack Overflow but also by most social media sites including Facebook, Twitter and Instagram. Tags and hashtags are very useful and can help with query-based like searches to retrieve information faster and more accurately. Especially, in the technical domain for "question and answer" sites such as Stack Overflow or Quora is it essential that the tags chosen during the post creation provide a correct representation of the submitted post.

When users submit questions on Stack Overflow they need to attach at least one and up to five tags to their question. These tags broadly identify the programming language talked about, the problem type in discussion and maybe some other fine-grained categories the question belongs to. The tags associated to each question help with information retrieval or user queries. For example, it may be very useful when users try to identify duplicate questions or related questions to a particular problem. Several approaches [2]–[5] for automatically generating tags from short questions or text have been recently developed. The Stack Overflow dataset is perfect for this problem as it provides the ground truth (as the author of the question adds these tags). The problem with this multi-label, multi-class classification approach is that it is hard to reach good accuracy.

In the last couple of years many deep learning algorithms were developed and provided better performance especially for computer vision problems like object classification in still images. Deep learning methods are considered special sub-category of machine learning methods, and part of the artificial intelligence field. They were developed based on the classical neural networks idea and are capable of learning data representations very well at multiple levels of abstractions, given the multiple layers of the network structure. Deep learning has been successfully applied to many fields as computer vision, speech recognition and translation, natural language processing and many others. For this project, deep learning can bring several benefits in learning natural language and code representations in order to improve the prediction of tags.

Collobert and Weston [6] were the first to use deep learning for six natural language processing (NLP) tasks that can be seen as tasks assigning labels to words. They applied multitask learning using a single convolutional neural network architecture for part-of-speech tagging and other several tasks, given input sentences. The feature extraction step is integrated into the deep learning architecture which improves the generalization performance. The different NLP problem of sentence to sentence learning was undertaken in [7] using three multi-task encoder-decoder networks: one-to-many setting, many-to-one setting, and many-to-many setting.

More recently, sentiment analysis for movie reviews tasks were successfully investigated by Liu et al. [8] using multiple recurrent neural network architectures with sharing information to model text with task-specific and shared layers.

Our main goal for this project is to use newly developed natural language and code representations together with deep learning algorithms to improve the prediction accuracy for automatically generating Stack Overflow question tags. The main benefits of this approach are 1) to automatically identify the keywords in short text and source code snippets and 2) to determine what the best NLP and code representations and deep learning architectures and parameters.

Research Questions and Hypotheses

For this project we will focus on the following research questions.

- To what degree do different text and code representations influence prediction results of tag extraction techniques?
- To what degree do the top-n keywords from our approach and the standard machine learning approaches match the keywords chosen by the author of the Stack Overflow question?
- What are the best deep learning architectures and hyperparameters setups that can be successfully used to make tag predictions?

Research Approach

The problem of tag prediction can be considered a multi-label classification problem. A deep learning algorithm will be trained using the question's title, sentences extracted from the question's text, and lines of code in order to predict the top five tags. The samples in the training set will be questions from an existing Stack Overflow dataset [9] extracted from the Stack Overflow data dump.

The first step in our approach will consist in representing the Stack Overflow posts as word embeddings. These representations are generally generated by creating a matrix of what words are similar in a large corpus. Several good pre-trained word embeddings have been recently developed. The most simple and popular model *word2vec*, also called the *skip-gram* model, learns word vectors using a neural network predictive model [10]. Another popular word vector representation is the *GloVe* [11] system creates a cooccurrence matrix and then reduces it using singular-value decomposition (SVD). In *Paragraph Vector* [12], an unsupervised method based also on SVD is used to learn fixed-length feature representations for documents. We will evaluate all the above methods in terms of their predicting results.

From source code chunks, two types of features can be extracted. First, there are the low-level lexical features such as "word length", "contains digit", "word unpredictability", "contains uppercase", "is all uppercase". Second, using code2vec [13] or python2vec to extract word embeddings from source code in a similar way with word2vec can help to improve the prediction.

The second step consists of running experiments using the word and code embeddings generated during the first step as inputs for multiple setups of convolutional neural networks (CNN) for the classification task. CNNs, which were originally invented for computer vision problems, build on layers for convolving filters and have proved to improve on other methods when it comes to solve document classification problems [14]. However, these models require researchers to specify an exact model architecture and set precise values for hyperparameters,

including the filter region size and the regularization parameters. This can be accomplished only by designing and running multiple sets of experiments.

Overall these techniques provide several different ways for short texts and programming codes to be processed and classified based on a set of training data. To train and test these models, the dataset will be divided into a training set 60% and a smaller testing set 40%. From each question we will randomly choose five positive tag labels and five negative tag labels. We will also compare the results obtained with CNN with results from classical classification algorithms such as: linear support vector machines (SVM), Naïve Bayes, and random forests.

Evaluation Plan and Expected Outcomes

Our models will be evaluated using accuracy, precision, recall, mean F1-measure and receiver operating characteristic (ROC) curves. The ROC curves will be obtained by computing the true positive rates versus the false positive rate. We plan to test several models using different sets of representations. All these measures will be used in conjunction to obtain a more comprehensive picture of the model's capabilities.

The students will report their findings in technical research papers and posters. Our team will likely submit these findings at regional and/or national scientific meetings like Grace Hopper Celebration, ACM Richard Tapia Celebration of Diversity in Computing, IEEE International Conference on Machine Learning and Applications and Mining Software Repositories Conference. They will also participate in the YSU STEM Showcase that is held every year and open to all nearby schools. This will help get more students interested in Computer Science projects. Results generated by this project will be included in one or more future manuscripts, with participating students afforded full opportunity to share the responsibilities and rewards of authorship.

Research Team

The goal of the CREU project is to encourage females and minorities to pursue graduate work and study in the field of computer science. This project will provide realistic research experience for our female undergraduates, by active involvement in the planning, execution and interpretation of scientific research. Well-developed research projects can significantly enrich the educational experience for undergraduate students. Working on this research project, the students will be able to enhance their programming and data analytics skills, apply those skills to investigate scientific problems, learn how to formulate questions and problems and to participate in the discovery of new knowledge. A good research experience can foster an enthusiasm for lifelong learning and a desire to continue education beyond the bachelor's degree. The students will be exposed to both sides of the scientific investigation: hypothesis testing and development of theoretical explanations of observations. No science education is complete without research related activities, technical writing and oral presentations.

Hannah Senediak is a junior student pursuing a degree in Computer Science. She started as an electrical engineering major but changed to Computer Science at the beginning of her sophomore year. She has had previous experience being a part of other team research projects through Choose Ohio First Scholarship and in a Physics lab on campus. Based on her background, Hannah will probably assume a student leadership role in our team.

Alexandra Ballow is a sophomore student, math and physics major with a minor in computer science. She is very excited to get involved in a computer science research project and see how it compares to the math projects she has already done. She is a member of Choose Ohio First so

she values research and scholarship. As a female in male dominated fields she is hoping this research project will give her the confidence to excel along with the men.

The girls participating in next year's CREU have the knowledge and skills that work synergistically to achieve the goals of this project. Two of them have been part of the Choose Ohio First Scholarship program at YSU and will continue with the program next year. They have been working on a research project, that resulted in a poster presentation during the 2017 – 2018 academic year. This proves that our students are motivated and dedicated to complete their work even under pressure. The students understand the 10 hours/week commitment requirement and a willing to dedicate this time to the project. We will encourage all our students to become members of the ACM-W student chapter at YSU and strive to build awareness by being change leaders in STEM fields. The students intend to present this project at OCWiC 2019 Ohio Celebration of Women in Computing and also at the next year's QUEST at Youngstown State University, a program highlighting student research.

Student Activities and Responsibilities

Specific tasks for the students will include: literature search and review, review write-ups, reading, presenting, and discussing research articles, designing and implementing the experiments, data processing, data analysis and interpretation, running data mining algorithms, summarizing and preparing results for presentations and publications, OCWiC 2019 and YSU QUEST 2019 participation and writing the final report. The students will also be mentored to prepare conference papers and posters to be submitted to other research conferences.

The primary responsibility of the students is to participate in all phases of the project: proposal, development, experiments, and dissemination. The students will be required to do weekly independent work and to schedule team meetings with the faculty advisors. The faculty advisors will meet with the students every week. Email and a central repository will be used for questions, announcements, and document interchange. A blog will also be setup during the first week with clear shared responsibility for updates. We will use basecamp to share a to-do list and keep track of all the tasks required to the success of the project.

Faculty Activities and Responsibilities

Dr. Alina Lazar will work to actively mentor the students and continuously supervise their progress during the one-year period. Together with her colleague, they will meet with the team on regular basis to guide their activities and answer students' questions related to the project. Dr. Lazar has extensive experience in data science, machine learning and artificial intelligence and she has written several papers related to these fields. Dr. Lazar will help the students during all steps of the project, from literature review, preparing the data, and running the experiments to data analysis, writing the final report and preparing conference papers. She will mostly lead and oversee the machine learning aspects of the project.

Dr. Bonita Sharif has conducted many empirical studies in the field of software engineering using various means of data collection such as questionnaires and biometrics such as eye tracking hardware and software. She will be responsible for developing concrete research hypotheses for each of the research questions mentioned above. She will mentor the students to design the experiments that address the research questions. This will involve experimental design including selection of variables, selection of tasks, and appropriate methods before the study is implemented. The students will be part of the design and variable selection. After the

data is collected, she will also be responsible for guiding the student to choose appropriate methods of analyzing the data and help with publication write-up.

The overall guidance and mentoring will not refer only to this project but also provide insights about how to apply and how to succeed in graduate school, about being a female scientist and what options are available after graduate school.

Project Timeline

Please refer to Table 1 for the project timeline.

Table 1. Project Timeline

Task Name	2018				2019							
	S	O	N	D	J	F	M	A	M	J	J	A
Literature review on tag prediction for Stack Overflow documents.												
Clean, format and preprocess the data for prediction. Run an initial data analysis visualization step.												
Generate several word embeddings.												
Develop and conduct experiments on Stack Overflow questions using the multiple setups of CNNs.												
Statistically analyze the classification results												
Dissemination of results through papers and communications at specific conferences.												

Appropriateness for CREU funding

Dr. Lazar and Dr. Sharif have been working on CREU projects since 2012. Dr. Lazar first mentor a student with CREU funding in 2004. We have been doing this long enough to see how important the CREU funding is to encourage women to pursue graduate studies. We have already seen some of our students finished their Ph.D. It is our belief that the CREU opportunity is very important.

For the proposed project we are requesting $3000 for each student. The additional $1500 will be used to support the student travel to Grace Hopper celebration, Tapia or other research conferences. While working on the project at least two students will also receive Choose Ohio First Scholarships and other department's scholarships. Students will be encouraged to apply for the Undergraduate Student Research Grant Award sponsored by the Youngstown State University and other scholarships. We plan on applying for the summer extension next February. A summary of the budget is given below.

Items	
Academic year for Hannah	$3,000
Academic year for Alexandra	$3,000
Travel allowance	$1,500
Total requested:	**$7,500**

Role of the CREU project within the larger scope of this research

The Department of Computer Science and Information Systems at Youngstown State University has always valued undergraduate research and provided resources for projects. The chair, the dean, and the provost all encourage research activity and provide faculty with necessary time to conduct such research with students. The faculty advisors created a new usability lab in 2012 as part of an internal grant that has access to state-of-the-art machines and eye trackers that students

can use for research projects. Another lab that our students will use is the Software Engineering Research and Empirical Studies Lab that was established by Dr. Sharif. The students will use both labs for data analysis and running the experiments.

The results of this project will directly impact other projects that are currently underway at the Software Engineering and Empirical Studies Lab. The girls will occasionally collaborate with other lab members to make sure the work is extensible to other related projects. This will help advance not just this project but also other projects that are conducted in the lab.

Prior results of CREU projects

Bonita Sharif and Alina Lazar worked with student Natalie Halavick during the CREU 2017-2018 on the project titled: "*An Exploratory Study on the Information Seeking Behavior of Developers on Stack Overflow*". The goal of the project was to study how software developers seek for information on Stack Overflow. We learned via an eye-tracking study how developers find information based on the task they are given to solve. Natalie presented the results of this project at YSU's QUEST in April 2018. We also submitted a poster for the Grace Hopper Celebration GHC 2018. If accepted Natalie will be presenting this work in September 2018.

Alina Lazar and Bonita Sharif worked with students Jenna Wise, Ali Morris, and Alyssa Pawluk during the CREU 2016-2017 project titled: "*Improving Stack Overflow Tag Prediction Using Eye Tracking*". Jenna presented the results of this project at YSU's QUEST in April 2017. We also presented the results of this work at the Ohio Celebration of Women in Computing (OCWIC) in February 2017 and at the Choose Ohio First Research Poster Conference in April 2017. Jenna won best poster at the Choose Ohio First Conference. In addition, Jenna presented this as a poster submission at the upcoming GHC in 2017. Jenna is currently in the PhD software engineering program at CMU.

Bonita Sharif and Alina Lazar worked with students Jenna Wise and Jessica Whitely during the CREU 2015-2016 project titled: "*Predicting Areas of Interest in Code Reading*". Jenna presented the results of this project at YSU's QUEST in April 2016. We also got a poster accepted at ACM TAPIA Conference 2016. Jenna will be presenting this work in September 2016.

Bonita Sharif and Alina Lazar also worked with students Jenna Wise and Jessica Whitely during the CREU 2014-2015 project titled: "*Mining Eye-tracking Data to Determine Developer Expertise and Task Difficulty in Software Development*". Jenna and Jessica presented the project "*An Eye-tracking Experiment Studying Problem Solving Behavior*" at the Ohio Celebration of Women in Computing (OCWIC 2015) conference and a talk at YSU's QUEST 2015 - a forum for student scholarship. Jessica also presented her senior project *"Towards Understanding Student Problem Solving Behavior in Algorithms via Eye Tracking"* in the department.

Alina Lazar and Bonita Sharif advised Sarah Ritchey during the CREU 2013-2014 project titled: "*Classification Algorithms for Detecting Duplicate Bug Reports in Large Open Repositories*". We got two papers published at the International Conference on Mining Software Repositories in May 2014. Sarah also presented this work at YSU's QUEST 2014, a forum for student scholarship, Ohio MAA Spring Sectional Mathematics Meeting, and the Pi Mu Epsilon meeting. We also got a poster accepted at the Grace Hopper Celebration 2014. This was presented by Dr. Sharif as Sarah had other engagements during that time. Sarah is now doing her Ph.D. at Duke University.

Bonita Sharif and Alina Lazar guided student Rachel Turner during the CREU 2012 -2013 project titled: "*An Empirical Investigation on Code Debugging and Understanding: An Eye-*

tracking Perspective". Rachel presented the poster "*C ++ vs. Python: An Eye-tracking Assessment*" at the Ohio Celebration of Women in Computing (OCWIC 2013) conference and a talk at the YSU's QUEST 2013 A Forum for Student Scholarship. We also published a paper at the Eye Tracking Research and Applications Conference (ETRA) in March of 2014.

Alina Lazar advised two prior CREU projects in 2004-2005 and 2006-2007, and one MROW project in 2007-2008. Darcy Davis the participant student in the 2004-2005 project just finished her PhD in Computer Science and Engineering at Notre Dame University. Louise Popio who participated as a student during 2006-2007, received her Master's in Information Sciences and Technology from Pennsylvania State University in 2010. Irena Lanc participated in the 2007-2008 MROW project and received her PhD in Computer Science and Engineering from The University of Notre Dame in 2014. Erin Pfeil the other student that did the 2007-2008 MROW project, is working on her PhD in Ecology at the University of Pittsburgh.

References

[1] A. Ahmad, C. Feng, S. Ge, and A. Yousif, "A survey on mining stack overflow: question and answering (Q&A) community," *Data Technol. Appl.*, vol. 52, no. 2, pp. 190–247, 2018.

[2] A. K. Saha, R. K. Saha, and K. A. Schneider, "A discriminative model approach for suggesting tags automatically for stack overflow questions," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, 2013, pp. 73–76.

[3] M. Lipczak and E. Milios, "Learning in efficient tag recommendation," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 167–174.

[4] C. Stanley and M. D. Byrne, "Predicting tags for stackoverflow posts," in *Proceedings of ICCM*, 2013, vol. 2013.

[5] S. Beyer and M. Pinzger, "Synonym suggestion for tags on stack overflow," in *Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension*, 2015, pp. 94–103.

[6] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[7] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," *ArXiv Prepr. ArXiv151106114*, 2015.

[8] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *ArXiv Prepr. ArXiv160505101*, 2016.

[9] "Facebook Recruiting III - Keyword Extraction," *Facebook Recruiting III - Keyword Extraction*. [Online]. Available: https://www.kaggle.com/c/facebook-recruiting-iii-keyword-extraction.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[11] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[12] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[13] U. Alon, M. Zilberstein, O. Levy, and E. Yahav, “code2vec: Learning Distributed Representations of Code,” *ArXiv Prepr. ArXiv180309473*, 2018.
[14] Y. Kim, “Convolutional neural networks for sentence classification,” *ArXiv Prepr. ArXiv14085882*, 2014.