Natural Language Processing for Understanding Novel Text Summarization and Media Bias

Emily Lo, YSP Student, Wellesley High School
Alyssa Shelburne, YSP Student, Westwood High School
Luyang Huang, M.S. Student, Boston University
Marshall White, B.S. Student, Northeastern University
Dr. Lu Wang, Khoury College of Computer Sciences, Northeastern University

Young Scholars Program
Northeastern University Khoury College of Computer Sciences
Northeastern University Center for STEM Education
The Social Science Research Council SSRC
Office of the Director of National Intelligence IARPA BE THE FUTURE
NSF

Abstract

The goal of Natural Language Processing (NLP) is to have computers understand human syntax and language[1]. This project aims to explore and design NLP and machine learning algorithms to solve novel text understanding problems. Specifically, we hope to develop automatic information extraction systems for summarization and usage on media bias and fake news analysis. Outlets with different ideologies were studied based on word usage and semantic structure to develop a framework for extraction methods.

Background

NLP is the intersection of artificial intelligence and linguistics. Applications of NLP include voice recognition, media bias, and text summarization[1]. NLP is not exclusive to science. Whether politics, business, or journalism, NLP can be applied to anything involving spoken or written language.

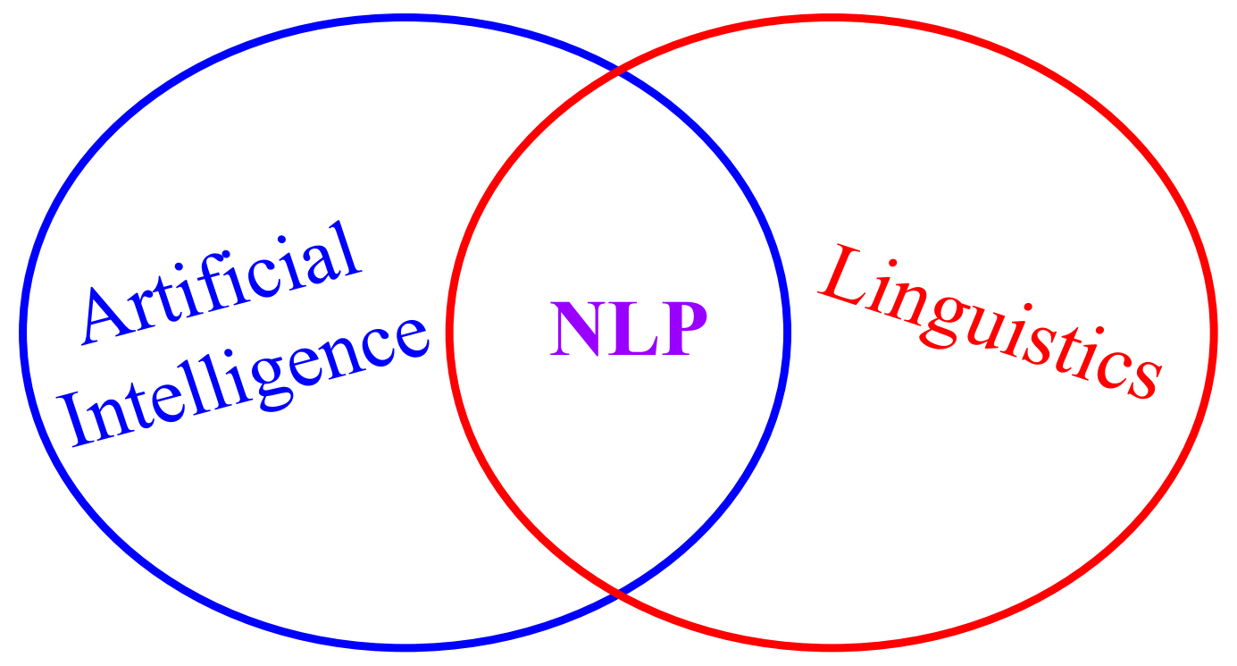


Figure 1. NLP helps computers to analyze, understand, alter, and generate human language.

Methods

Goals for Summarization Project: Use NLP to understand how to generate summaries of long texts best and reformat Congressional Report Services (CRS) documents.

We used Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to determine the quality of computer generated summaries by comparing them to human summaries[2].

Goal for Media Bias Project: Use NLP to manually detect and categorize bias in news articles.

"My toy broke" vs. "I broke my toy"

Figure 2. Passive voice can be utilized to shift blame[3].

We studied computational linguistics to identify frames and grammatical structures that are prone to implicit sentiment to create a framework for automatic bias detection algorithms.

Results and Discussion

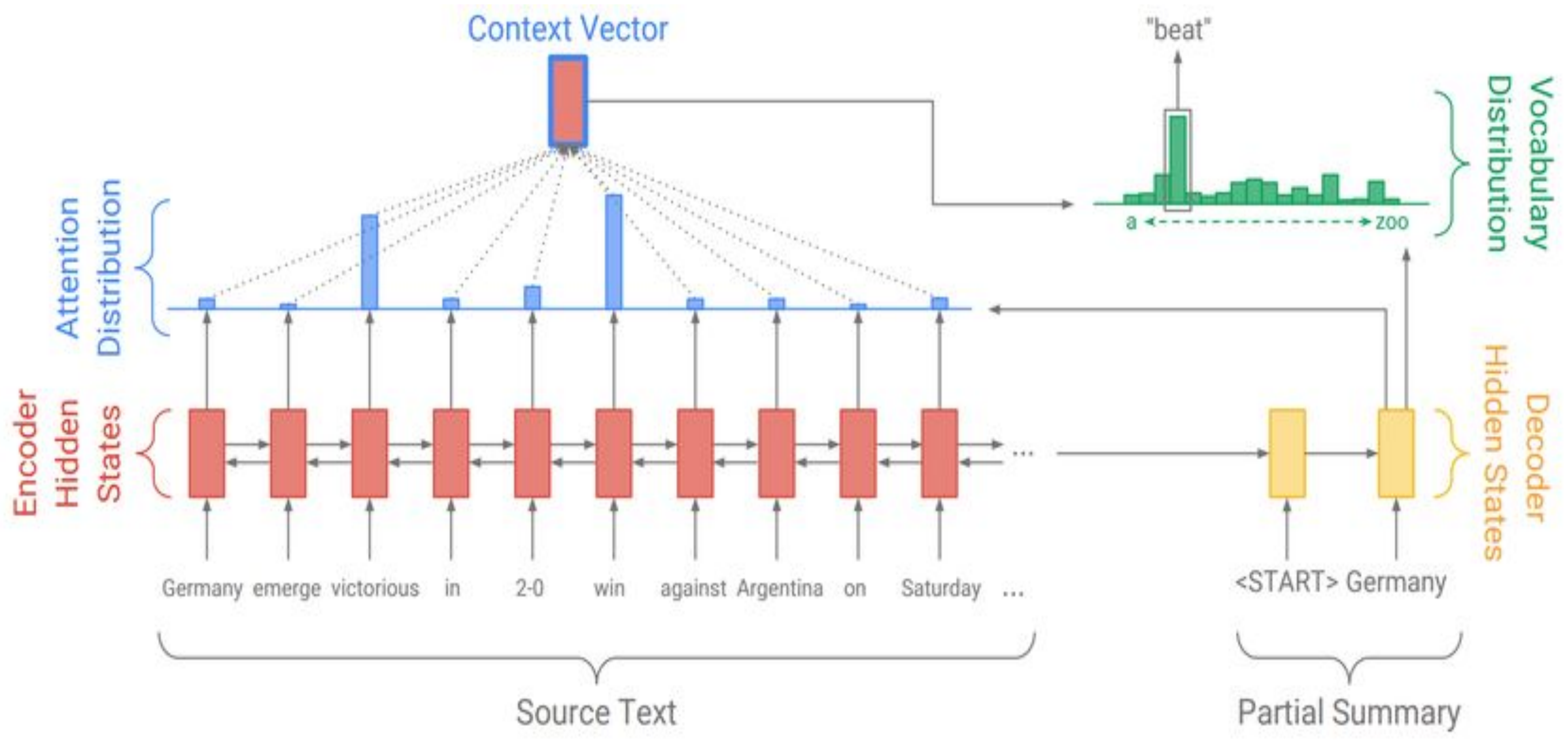


Figure 3. A sequence to sequence model takes words, and categorizes them by weight, and then tries to predict the next element of the sequence[4].

Original Text: The 30-year-old scored four goals in 22 appearances for the Spitfires last season and spent three months on loan at Aldershot Town. Lafayette has previously had spells with Welling , Woking and Luton Town. Dover host Boreham Wood in the national league on Tuesday , having begun the 2016-17 season with a 0-0 draw at Wrexham on Saturday. Find all the latest football transfers on our dedicated page.

Reference (human summary): National league side Dover athletic have signed Ross Lafayette following his departure from Eastleigh.

Lead 3 Summary (first three sentences)
ROUGE-1 F1 = 0.1333 ROUGE-2 F1 = 0.0285 ROUGE-L F1 = 0.0693

Figure 4. ROUGE scores are calculated based on word overlap, bigram overlap, and Longest Common Subsequence based statistics.

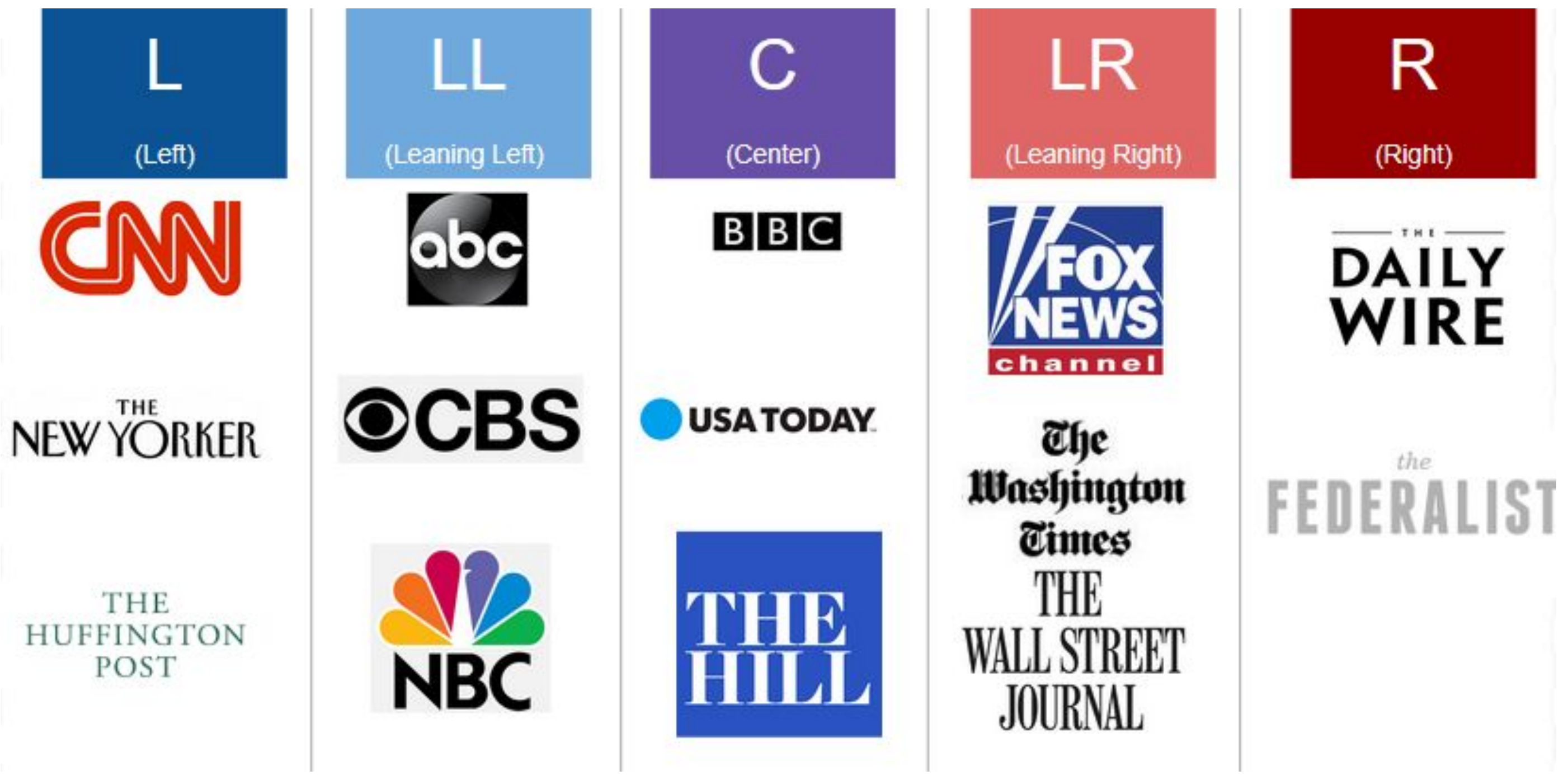


Figure 5. Information distributed by news outlets is not always reliable; some articles contain political bias that advocates wrong information[5].

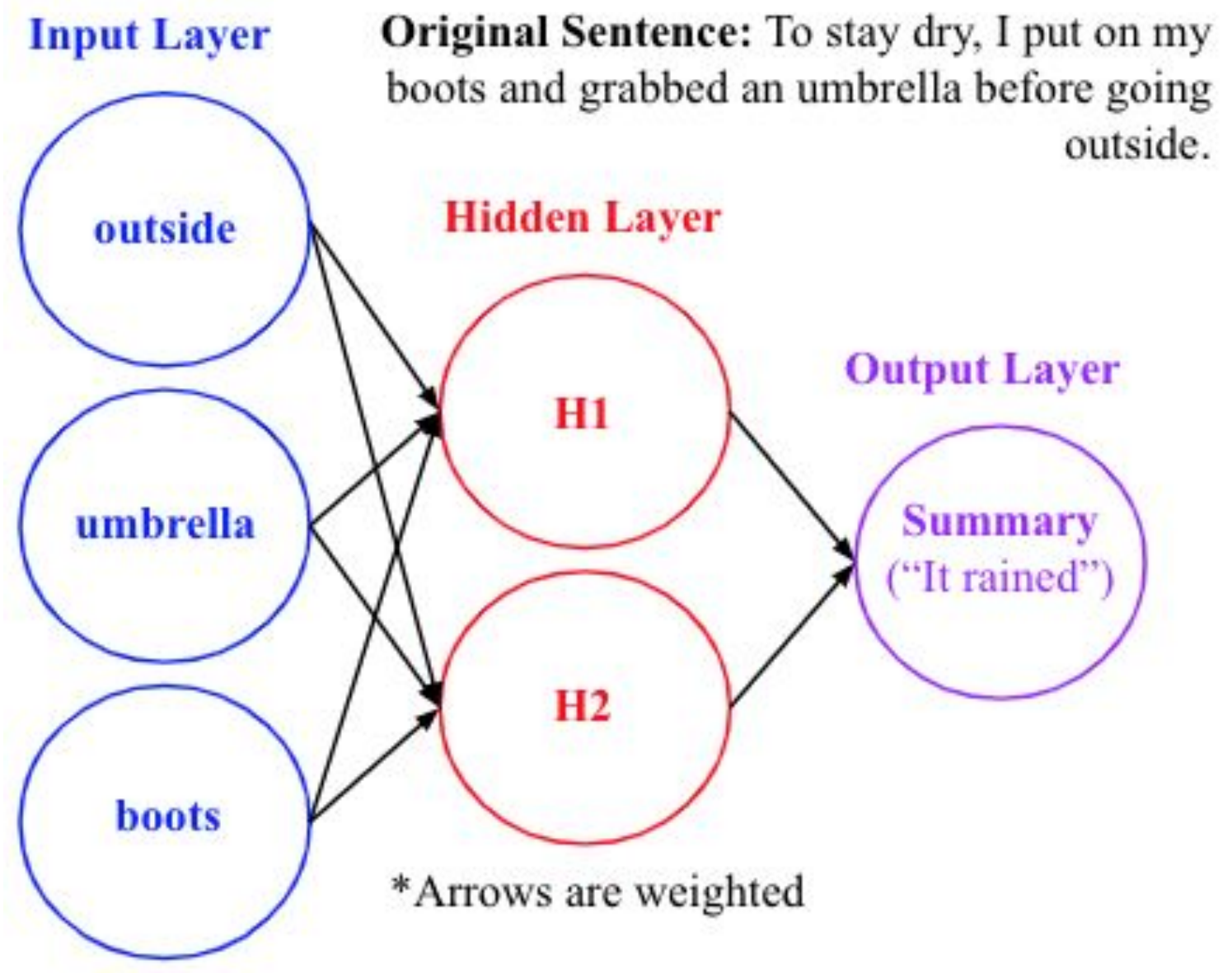
Summarized news articles help readers quickly obtain the main idea of an article without wasting time reading possibly unnecessary information. ROUGE helps indicate abstraction or lack of it in generated summaries. While many news outlets have a reputation for being biased, individual articles may not uphold the same idea.

Future Work

Summarization Project

238 cleaned CRS documents will be used to develop and test abstractive summarization methods that are capable of condensing thousand-page reports into several sentences.

Figure 6. Modeled after human brains, neural networks are algorithms designed to recognize patterns. They help computers to generate summaries that contain sentences not found in the original text[4].



Media Bias Project

The Lu Wang lab will continue investigating different bias strategies used by media outlets and create computational methods for detecting bias.

References

1. Kiser, M. (2018, January 02). Introduction to Natural Language Processing (NLP). Retrieved from https://blog.algorithmia.com/introduction-natural-language-processing-nlp/
2. Lin, C. (n.d.). ROUGE: A Package for Automatic Evaluation of Summaries. Retrieved from https://www.aclweb.org/anthology/W04-1013
3. Greene, S., & Resnik, P. (n.d.). More than Words: Syntactic Packaging and Implicit Sentiment. Retrieved from https://www.aclweb.org/anthology/N09-1057
4. See, A., Lu, P. J., & Manning, C. D. (n.d.). Get To The Point: Summarization with Pointer-Generator Networks. Retrieved from https://nlp.stanford.edu/pubs/see2017get.pdf
5. Media Bias Chart. (2019, June 04). Retrieved from https://www.allsides.com/media-bias/media-bias-chart

Acknowledgements

Center for STEM Education
Claire Duggan
YSP Coordinators
Emily Chernich
Salima Amiji
Khoury College of Computer Sciences
Professor Lu Wang
Luyang Huang & Marshall White

Young Scholars Program at Northeastern University
Claire Duggan, Program Director

This work was supported by the National Science Foundation
Grant: #181334