Machine Learning: Improving Named Entity Recognition in Healthcare and Business Sector

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Kevin Liu

Spring, 2022

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

Briana Morrison, Department of Computer Science

Machine Learning: Improving Named Entity Recognition in Healthcare and Business Sector

CS4991 Capstone Report, 2022
Kevin Liu
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
xl3muv@virginia.edu

Abstract
Qualtrics, a software company that provides customer experience management solutions, wishes to increase its presence in the healthcare and business sectors. Managing customers requires understanding them first, so Qualtrics needs to improve its natural language processing (NLP) capabilities—specifically, named entity recognition (NER). I combined machine learning models with rule-based matching to improve NER scores. Since ML models are unpredictable, I used the trial and error method to determine which ML and rule-based matching combination achieves the highest scores. Additionally, I used the spaCy library to train and test NLP models, and Wikidata Query Service to obtain larger lexicons. By the end of the internship, the optimal and highest scoring combination of ML model and rule-based matching was determined for both the healthcare and business sectors. The model should be further refined to improve the precision and recall scores, so it can be deployed in the production environment.

1. Introduction

Imagine you are a high-level product manager at Apple. Views and opinions of the newly launched iPhone 29 will play an important role in the company's growth trajectory this fiscal year. A minor defect has broken the water-resistant seal in many units, but no one in the company knows. You could scroll aimlessly on Amazon, looking for constructive negative reviews and comments, but when there are millions of other reviews, it would be like finding a needle in a haystack. You decide to employ Qualtrics to perform customer experience analysis. One of the reviews processed by Qualtrics's NLP engine states: "I dropped my Apple iPhone 29 next to the bank and now it doesn't work." If we use a rule-based NLP model, then it might mistake "Apple" as the fruit or "the bank" as a financial institution, therefore disregarding this important review.

This is where machine learning NLP models have an advantage over rule-based ones. ML models are able to identify and classify words by their context. Ideally, the ML model is able to identify and tag "Apple" as an organization and "bank" as geo-location. The identifying and classifying aspect of NLP is called Named Entity Recognition (NER), and the goal of my project at Qualtrics was to improve NER precision, recall, and f-1 scores in the business and healthcare sectors. Different sectors have different vocabulary, punctuation, and sentence structure, so another part of the project was also to determine if models can be created to function across both sectors.

2. Related Works

Many large technology companies, such as Google and Amazon, use machine learningbased NLP models. The technical information regarding these models is difficult to find since they are valuable and proprietary resources. There are, however, some publicly-available models trained for the healthcare industry. One such package is scispacy, created by AI2. According to AI2, "scispaCy is a Python package containing spaCy models for processing biomedical, scientific or clinical text." (allenai, 2022) The models are either pre-trained full pipeline models or NER pipe only models. As part of my project, I had to integrate the NER component with existing components that my team created previously for Qualtrics. I could not simply use models in scispacy since I could not integrate them. However, I used the BioCreative V CDR corpus on which scispacy's en_ner_bc5cdr_md model was trained. Using the BioCreative V CDR corpus, I was able to train and integrate a NER model.

I also used the default en_core_web_lg model from the spaCy library (Explosion 2022). en_core_web_lg is a full pipeline model trained for general purpose corpus. It is not specialized for any particular type of corpus, so I can use it as a base model and train it further to be specialized.

3. Process Design

Because this internship project at Qualtrics is one large research spike, there was no clearly defined plan for it. The strategy of the project relied on iterative discovery. Using the information gathered in the previous iteration, I created my plan for the next iteration. The high-level instruction I received from my manager was to create and test models trained on different combinations of training datasets and lexicon dictionaries.

3.1 Model Evaluation

The models are evaluated on their precision, recall, and f-1 scores; or more specifically, on the model's overall precision, recall, and f-1 scores and the scores on each individual tag. For healthcare corpora, there are two tags: Chemical and Disease. For business corpora, there are three tags: Organization, Product, and Person.

To further understand what these scores mean, examine the diagram below.

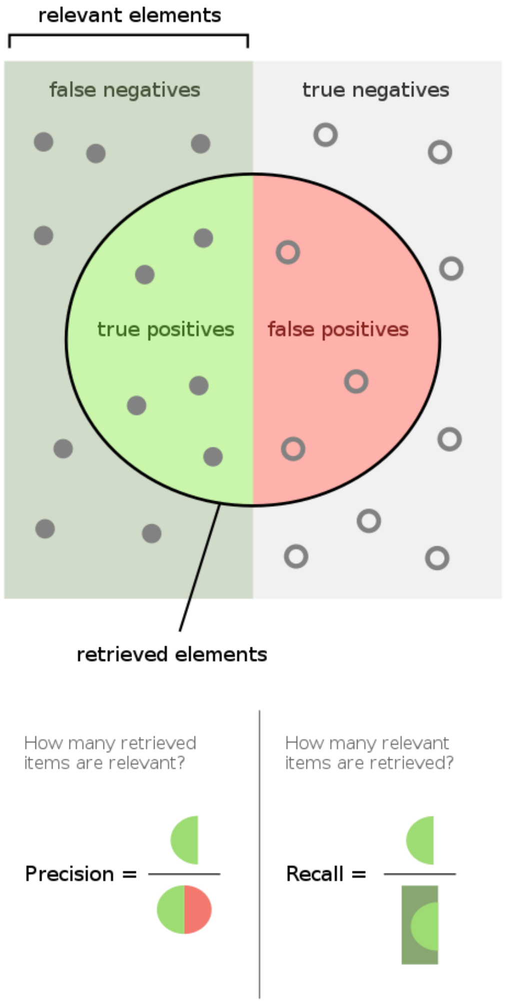

Figure 1: Precision and Recall Diagram. [3]

Figure 1 shows that when false positives decrease and false negatives increase, precision increases. When false positives increase and false negatives decrease, recall increases. The f-1 score is the harmonic mean of precision and recall.

3.2 Project Procedure

I explored numerous combinations and went down multiple paths during the spike project. Some of the models created did not improve scores but provided valuable insights for improvements on the following models. First, I created a rule-based matching healthcare model. This particular model only had the lexicon pipe implemented. The lexicon pipe is created from lexicon lists that already exist in the company's database. Next, I improved this rule-based approach by querying lexicons from Wikidata. The Wikidata queries effectively expanded the lexicon pipe size. These two models generally tested the limits of the rule-based matching approach and provided bases for measuring some later models.

Since these two healthcare models only comprised a lexicon pipe, I was able to isolate the lexicon component and roughly perfect it for later reuse. This task was made easier due to the isolation of the component which allowed me to deal with just one variable instead of multiple variables in a multi-piped model.

After rule-based healthcare models, I worked on ML healthcare models. I created three ML NER healthcare models in total. All of them have the same ML NER pipe but differ in the size of their lexicon pipe. One model has no lexicon pipe, one model has the lexicon pipe created from company data, and the other model has the lexicon pipe created from Wikidata queries. The lexicon pipe is appended after the NER pipe, and the tags identified by both of the pipes are unioned together. The NER pipes of these models are the same, and they are trained from the BioCreative V CDR dataset. I performed hyperparameter tuning on all ML models by changing parameters such as dropout rate, patience, seed, network width, etc.

After healthcare models, I began working on business models. The creation of business models follows a very similar process to the healthcare models. First, a rule-based matching approach is taken to create the lexicon pipes. Next, I created ML NER pipes from the two sets of business data given to me. One of the datasets belongs to Qualtrics while the other belongs to Clarabridge, a recent acquiree. I created three models for both sets of data. Similarly to the healthcare approach, one without lexicon pipe, one with lexicon created from company data, and one with lexicon created from Wikidata queried data. Next, I trained a separate NER pipe from the combined training data of Qualtrics and Clarabridge and created three more models like previously. In total, I created nine ML business models and two rule-based matching models.

Last, I attempted to create a functional model for both healthcare and business corpus. I trained a NER pipe on the combined datasets of BioCreative V CDR, Qualtrics, and Clarabridge. I then tried various lexicon pipes to determine which one scored the highest.

3.3 Challenges

I faced two major challenges during this project. The first obstacle appeared when I trained ML NER pipes for healthcare models. Corpus from the healthcare sector uses unconventional punctuation. For example, most corpora contain dashes in chemical names, symbols to represent units, etc. The standard tokenizer from spaCy's en_core_web_lg model cannot tokenize healthcare corpora properly, which led to problems during training. To solve this issue, I created a custom tokenizer designed specifically to parse and tokenize healthcare corpus.

The second obstacle I faced was combining Qualtrics and Clarabridge business datasets for training and testing. The two datasets use different labels. For example, one uses “Brand” while the other uses “Organization” to represent the same entity. To solve this issue, I unified the entity tag names across both datasets and prepared them for training.

4. Results

Across the span of the project, I created around 15 models. I saw significant improvements to recall and f-1 score in the ML healthcare models when increasing the lexicon pipe size. In other words, for healthcare corpora, the increase in recall and f-1 scores is correlated with the size of the lexicon pipe. For the business models, although I observed an increase in recall scores corresponding to increasing lexicon pipe sizes, those recall scores cannot be justified by the decrease in precision and f-1 scores. The decrease in f-1 scores confirms the minimal increase in recall does not offset the drastic decrease in precision.

Another part of the project was to discover if it is possible to create a model that functions across both sectors. The result of the experiment is promising. The models trained from combined healthcare and business data performed no worse than the individual sector models. For certain tags, the combined model scored even higher.

In conclusion, I determined that lexicon pipe improves NER scores in the healthcare sector and not the business sector. In addition, it is possible to create a single NER model for both sectors.

5. Conclusion

Many think that customer experience management with NLP technology will dominate and revolutionize how management is done. It is a rapidly growing market with promising potential. Improving NER scores in the healthcare and business sectors is a small step in enhancing NLP, but it is an essential step. As NER performance is improved, Qualtrics can process customer feedback and data more accurately; therefore, customer opinions can be more correctly represented by computer data. This has huge potential such as saving a company from bankruptcy or assisting a company to win against its competitors. The value of an accurate and efficient NLP engine is immense which is why they are often proprietary, and this project contributes a small and influential part.

6. Future Work

For the future of the project, we wish to further refine the most optimal performing model which is the combined healthcare and business model with lexicon pipe. In essence, we need to utilize the information discovered during this research spike to create production-level models. To reach the production standard, precision and recall scores will need to be further optimized by methods such as hyperparameter tuning. After creating a production-level model, we want to integrate it with the existing NLP engine at Qualtrics.

Additionally, we also want to consider training models from other available datasets and compare the results with the models created in this project. If models trained from another dataset show better results, then we might consider utilizing those models instead of the ones trained in this project.

References

[1] allenai. Retrieved September 21, 2022 from https://allenai.github.io/scispacy/

[2] Explosion. Explosion/spacy-models: models for the Spacy Natural Language Processing (NLP) library. Retrieved September 21, 2022 from https://github.com/explosion/spacy-models

[3] Wikimedia Foundation. (2022, September 25). *Precision and recall*. Wikipedia. Retrieved October 20, 2022, from https://en.wikipedia.org/wiki/Precision_and_recall