Rule-based Symptom Detection Task (BMI 550 Assignment 1)

Swati Rajwal
Emory University, Atlanta, Georgia, USA

Abstract
This report describes a rule based symptom detection system for a given set of Reddit posts. Based on the lecture topics covered so far, exact and inexact string-matching are used to identify the symptoms from the given dictionary. Experimental results showcase that exact string-matching approach achieved an overall F1-score of 0.63 compared to 0.50 of by inexact string-matching. Using the better performing system, the unlabeled dataset has been annotated.

Introduction

In NLP, rule-based tasks involve predefined pattern-based rules to identify text elements. They rely on human-defined rules instead of machine learning. Rule-based methods are effective for well-understood patterns. This assignment aims to develop a system for processing Reddit posts. It generates a file with post IDs, symptom expressions, CUIs, and negation flags (1 for negated symptoms, 0 otherwise). This file is used to evaluate the system's F1-score.

Methods

Manual annotations

Fifteen annotators manually annotated reddit posts for symptoms using an Excel file and a symptom dictionary. Each annotator completed four columns: (1) Symptom Expressions- symptoms separated by $$$, (2) Standard Symptom- expressions from the dictionary, (3) CUI for each symptom, and (4) Negation Flag - 1s indicating negated symptoms.

Preprocessing

After removing the duplicates there were 153 unique texts in the dataset. Next, several preprocessing steps were performed: (1) Replacing newline characters ("\n") with spaces, (2) converting the entire text corpus to lowercase for uniformity ("Word", “WORD” and "word" are treated as identical), (3) removing specific punctuation marks. Also, apostrophes were removed from both the dataset and the list of negative words. For example, 'don't' became 'dont,' 'couldn't' became 'couldnt,' and so forth.

Rule based System

Two techniques are employed here: exact and Inexact string matching**.** The first method matches the precise symptom expression within the given text. For eg., in “I have fever”, the symptom “fever” matches, while “fevers” does not. The regular expression, *r'\b'+symptom+r'\b'*, is employed for exact matching. There are limitations to exact string matching such as its inability to detect spelling errors like 'fever' written as 'feber'. Alternatively, levenshtein ratio (minimum number of single character edits required to change one word to another) combined with thresholding and windowing is a useful mechanism for detecting inexact matching. The system also includes a mechanism to checks if a symptom is negative. It tokenizes[1] the text after the negation term is encountered, examines the first three words (or fewer) for a symptom match, and considers full-stop and nearby negations to determine if the symptom is negated.

Results

Annotation Agreement

The Inter-annotator agreement score (IAA) is evaluated by comparing my annotations with those of the rest of the class. Through manual inspection, it was found that there were minimal disagreements among annotators, primarily arising from the subjective nature of tagging 'other' symptoms. Therefore, these discrepancies are potentially due to personal judgment than clear errors. In all remaining instances, the average IAA score turns out to be **0.926** (see Appendix, Table 1), indicating a higher level of agreement among annotators.

System Evaluations

The system generated one exact string-matching based annotation file which was evaluated against the gold standard via an evaluation script. Various metrics such as recall, precision and F1-score are evaluated, and the trend can be seen in Figure 1 (or Appendix Table 2). For various threshold values, a number of inexact string matching-based annotations were generated, and the results are as shown in Figure 1. For threshold value of 0.9, the inexact system achieved maximum F1 score (0.50) which is still less than F1 score by exact matching system. The code, resultant files and various other plots can be found in GitHub repository.

[1] https://www.nltk.org/api/nltk.tokenize.html

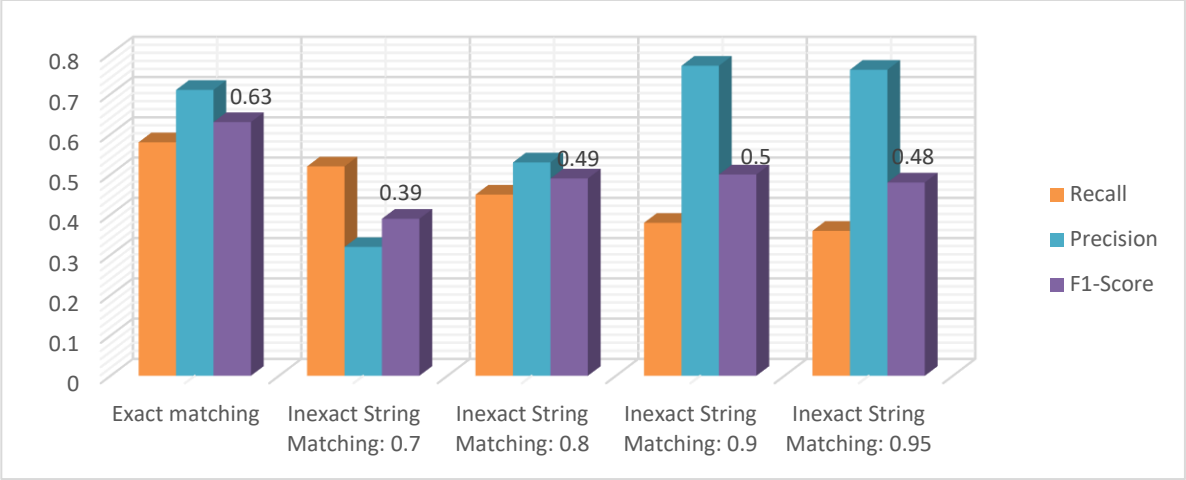

Figure 1. Results on given dataset against gold standard.

Error Analysis

After running the evaluation script between gold standard and rule-based system annotations, various sources of errors were identified by manually analyzing the wrong annotations. There were occasions when the rule-based system failed to identify a symptom mainly because the symptom did not precisely match symptom string provided in the dictionary. For post id 'g025b1' the system could not identify and classify "dull pain on my eye" as a symptom. This is the limitation of a rule-based system. Other errors of interest were also identified. For instance, for post id 'hiakr5', the system could not accurately identify the symptom 'cough' as negative in the sentence '*very little cough that's gone away*'. The reason for this could be due to the fact that the list of negations provided does not encompass all possible negative words and phrases. Some errors stemmed from issues within the gold standard itself, as outlined in the Appendix. Posts with id 'h8wb3q', 'g7vgeq', 'ho666d', 'gwojzx' were wrongly labeled FN for certain symptoms because the gold standard provided had problems. Surprisingly, for post id 'i3blx9' the system correctly predicted negative symptom, but gold standard mentioned that it was positive. For post id 'g025b', the gold standard failed to identify 'cough' as a symptom, while the rule-based system successfully detected it. As a result, this discrepancy resulted in a mismatch that was categorized as a FP by the evaluation script.

Symptom Distributions

The current rule-based system was executed on a collection of unannotated posts contained in the file unlabeledset.xlsx, and the results can be located here. According to the annotations generated by the system, we got the distribution of symptoms as shown in Figure 2. It can be observed that 'body ache & pain' is the most frequent symptom followed by 'Pyrexia', 'cough', and various others.

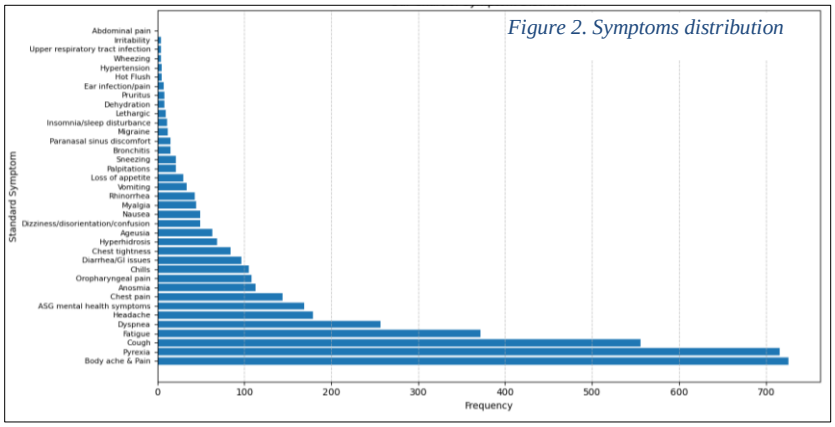

Figure 2. Symptoms distribution

Discussion and Conclusions

In this report, a rule-based symptom detection mechanism has been implemented and evaluated against the gold standard dataset. Two mechanisms namely exact and inexact matching have been utilized with exact matching (regular expression) resulting in higher F1-score compared to the other technique. In future work, specific machine learning tasks may be used in conjunction with rule based to improve the system's overall performance.

References

1. Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11(2):141-150. doi:10.1197/jamia.M1356.
2. Sarker A, Lakamana S, et al. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc. 2020 Aug 1;27(8):1310-1315. doi: 10.1093/jamia/ocaa116. ;PMCID: PMC7337747.

Appendix

Table 1. IAA score between my annotations and the annotations from other annotators was computed.

My file	**s15.xlsx**												
Other annotators	s4	s11	s8	s9	s10	s5	s2	s3	s1	S14	s6	s13	s12
# common instances	5	1	5	8	8	1	1	4	0	5	1	1	1
IAA (Cohen's kappa)	0.93	1	0.82	0.80	0.84	1	1	0.81	nan	0.83	1	1	1

Table 2. System annotated file evaluation results against Gold standard.

String Matching Technique Used	**Threshold value**	**Recall**	**Precision**	**F1-Score**
Exact matching	N/A	0.58	**0.71**	**0.63**
Inexact String Matching	0.70	0.52	0.32	0.39
	0.80	0.45	0.53	0.49
	0.90	0.38	0.77	**0.50**
	0.95	0.36	0.76	0.48

Table 3. Issues identified in Gold Standard

Post ID	**Excel sheet row number**	**Problem in Excel File**
h8wb3q	12	Wrong CUIs mentioned
g7vgeq	13	Wrong CUIs mentioned
ho666d	21	Extra $ symbols in CUIs column
gwojzx	28	Missing $ in CUIs list

Amendments made:

1. The gold standard evaluation file had data till row 35 of excel. But there was one issue on row number 376 of the excel file share. This row had nothing except an ‘s’ in the ‘TEXT’ column. Therefore, I decided to drop this row.
2. In the unlabeledset.xlsx file, row number 275 or post id ‘ga7gvk’ did not have anything in the ‘TEXT’ column. So I decided to drop that row.
3. When there is no symptom match in the given text, I am using $$$$$$ to denote empty value