Determining the Statistical Significance of Rules for Rule-based Knowledge-extraction Algorithms

Sushain Pandit, Krishnakumar Sridharan, Sateesh Kodavali

December 7, 2009

Abstract

Domain specific knowledge bases are often built from domain-specific texts using rule-based knowledge-retrieval algorithms. These algorithms are based on semantic extraction rules that process text using a parser, looking at the resulting parse trees & dependency graphs and then applying those rules to identify possible constructs for triple extraction. The performance of such algorithms critically depends on how capable these rules are in extracting the knowledge (in the form of triples) as a fraction of the total knowledge present in the text fragment. In this paper, we propose a way to statistically analyze the signicance of these rules based on the fraction of knowledge that they extract out of given text corpora.

1 Introduction

Building knowledge bases from domain-specific texts is a classic problem in knowledge acquisition. These knowledge bases are often built by extracting individual knowledge constructs from domain-specific texts using rule-based knowledge-retrieval algorithms [2]. These algorithms are based on semantic extraction rules that process text using a parser, looking at the resulting parse trees and/or dependency graphs and then applying those rules to identify possible constructs for knowledge extraction extraction. Part of this process is demonstrated in figure 1, which shows a dependency graph for the sentence - "Heart attack causes reduced average lifespan".

A possible extraction rule for this graph can be - "If only the labels $\{nsubj, dobj\}$ occur along a path in the graph, extract that path as a knowledge contruct". This rule alone would result in the extraction of - "*Causes-reduced-lifespan*", which doesn't seem to be a valid construct. However, if we add another (higher priority) rule that says - "If any of the labels $\{nn\}$, $\{amod\}$ appear along a path (edge) between two nodes, merge them to generate a complex node", then in conjunction with the first rule, we will get - "*{Heart, attack, causes}-reduced-{lifespan, average}*, which certainy looks closer to what we want.

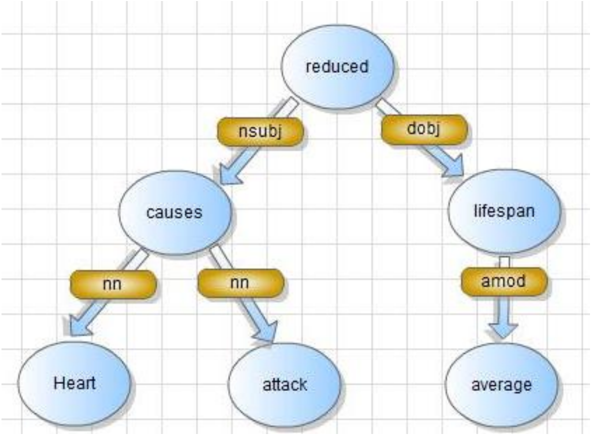


Figure 1: Example Dependency Graph

Thus, these rules critically determine the performance of an extraction algorithm that traverses such dependency graphs and uses these rules to extract knowledge structures. In the following sections, we formalize the notion of rules, the problem and proceed with our approach to statistically analyze the significance of these rules based on the performance of the algorithms that utilize them to extract knowledge.

2 Background and Motivation

Most of the existing rule-based algorithms are built with a specific domain of discourse in mind and the underlying rules are formed heuristically based on nature of the domain. As a result, an empirical need to evaluate the goodness of underlying rules doesn't arise. However, if one focuses on domain-independent knowledge retrieval systems, there is a need to have some measure for quantifying the goodness of the rules since different rules (or sets of rules) would lead to difference in the amount of knowledge extracted by the system on texts from different domains. This quantification of goodness can lead to interesting insights into the *kind* of rules that lead to better performance of the system, and this may ultimately lead to the characterization of an optimal (non-dominated w.r.t to the fraction of knowledge extracted) rule-set. There is fair amount of existing work on experimentally verifying the performance of rule-based algorithms that extract knowledge from domain-specific texts. Cartic et al [2] recently proposed rule-based algorithms for relationship extraction from text specifc to Biomedical domain and report the number of triples extracted for sample texts. There is also significant literature on rule-induction in text-based data mining, however our work is different in that we are not trying to induce rules but trying to test the statistical significance of existing rules. There is similar work for other domains, however, to the best of our knowledge, there isn't any published work describing a similar framework to measure goodness of rulesets for knowledge extraction and discover those rules that lead to the best generic performance across various text corpora.

3 Definitions and Notations

3.1 Terminology

We utilize the following notations in describing the approaches that we've implemented.

p_i : $i-th$ Condition or premise for a rule described below.
c_i :$i-th$ consequent corresponding to the $i-th$ premise.
r_i : A rule of the form, $\{p_i\} \longrightarrow \{c_i\}$ meaning $If\, \{p_i\}\, occurs,\, perform\, \{c_i\}$.
P : A labelled path in the dependency graph for a text sentence of the form show in figure 1 comprising of edges $\{e_i\}$.
$\{l_i\}$: An ordered set of labels on the edges $\{e_i\}$ along a path, P.

The rules for the text extraction utilized by the algorithms described are of the following form:

Definition 1: (*Extraction Rule)* For a path P, we define an extraction rule as,

$$r_i : \{\{P\, has\, an\, ordered\, set\, of\, labels\, \{l_{1i}\}\, along\, the\, edges\, \{e_{1i}\}\}, ...\} \longrightarrow \{Extract\, \{e_{1i}\}, ...\}$$

Here, r_i encodes the rule that if there exists a sequence of dependencies between the words of a given sentence s.t. that sequence is captured by the ordered set $\{l_i\}$, then the sequence of words forms a knowledge contruct and thus, the algorithm is recommended to extract it.

Definition 2: *(*Peformance ratio*)* We define the peformance ratio, P_k, due to the $k-th$ rule-set as the ratio of the number of extracted knowledge constructs to the total number of labelled constructs in text.

3.2 Problem Formulation

Given a set, $\mathbf{S} = \{R_k\}$, of rule-sets, $\{r_i\}$ and a corresponding set of performance ratios, $\{P_k\}$, our aim is to determine the significance of using a given subset T of $\mathbf{S}$ in an extraction algorithm. We intend to achieve this statistically by testing for the null hypothesis that the presence or absence of a rule/rule-set and performance ratio are independent, or in other words, there is not a significant difference in the performance ratio achieved by an algorithm when it uses T as against when it does not. We utilize a chi-square test of independence for achieving this as elaborated in the next section.

4 Approach

4.1 Chi-square Test for Determining Insignificant Set of Rules

Chi-squared test is a statistical hypothesis test in which the sampling distribution of the test statistic is chi-square, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough. The test is applied when we have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

We utilize this to test the null hypothesis that the presence or absence of a rule/rule-set and performance ratio are independent. We map our problem for testing the significance of a given subset T of $\mathbf{S}$ to a chi-square test scenario as follows.

We define the row variable as presense or absence of T. [meaning that the extraction algorithm uses the rule or it doesn't], and column variable as performance ratios divided into 4 classes: $C_1 : 0 \leq p < 0.2$, $C_2 : 0.2 \leq p < 0.5$, $C_3 : 0.5 \leq p < 0.8$, $C_4 : 0.8 \leq p \leq 1$. This is as indicated in the contigency table below.

	C_1	C_2	C_3	C_4	Row Totals
T	c_{T1}	c_{T2}	c_{T3}	c_{T4}	$\sum_{i=1}^{4} c_{Ti}$
$\sim T$	$c_{\sim T1}$	$c_{\sim T2}$	$c_{\sim T3}$	$c_{\sim T4}$	$\sum_{i=1}^{4} c_{\sim Ti}$
Column Totals	$c_{T1} + c_{\sim T1}$	$c_{T2} + c_{\sim T2}$	$c_{T3} + c_{\sim T3}$	$c_{T4} + c_{\sim T4}$	$Total\ Sum$

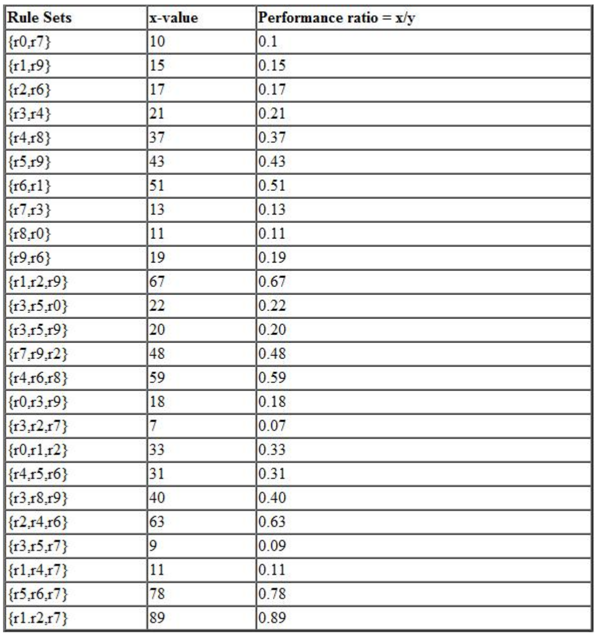
Here, c_{Ti} indicates the number of subsets of $\mathbf{S}$ which contain T such that the performance ratio for those subsets is within the class C_i. Similarly, $c_{\sim Ti}$ indicates the number of subsets of $\mathbf{S}$ which do not contain T such that the performance ratio for those subsets is within the class C_i.

Our null hypothesis, H_o is that the presence or absence of T and performance ratio are independent, or in other words, there is not a significant difference in the performance ratio achieved by an extraction algorithm when it uses T as against the case when it does not.

Once, we've this table, we calculate the test statistic, $\chi^2 = \sum_i [(O_i - E_i)^2 / E_i]$, where O_i are the observed cell-values in the table above and E_i are the expected counts calculated using $n_{i.} * n_{.j} / n_{..}$. Eventually, we calculate the p-value corresponding to this statistic. For our purposes, we set the significance level at 0.1 and if the p-value is less than this, we reject H_o.

<u>**Definition 3**</u> *(Rule Significance)*: We call a subset T of $\mathbf{S}$ to be significant *iff* the p-value of the test-statistic (as defined above) is ≤ 0.1.

Eventually, we intend to determine whether T remains significant across a multitude of text corpora since that would make it a domain independent generic rule, which can then be utilized in any extraction algorithm. Along these lines, we do our evaluations in the following section.



Rule Sets	x-value	Performance ratio = x/y
{r0,r7}	10	0.1
{r1,r9}	15	0.15
{r2,r6}	17	0.17
{r3,r4}	21	0.21
{r4,r8}	37	0.37
{r5,r9}	43	0.43
{r6,r1}	51	0.51
{r7,r3}	13	0.13
{r8,r0}	11	0.11
{r9,r6}	19	0.19
{r1,r2,r9}	67	0.67
{r3,r5,r0}	22	0.22
{r3,r5,r9}	20	0.20
{r7,r9,r2}	48	0.48
{r4,r6,r8}	59	0.59
{r0,r3,r9}	18	0.18
{r3,r2,r7}	7	0.07
{r0,r1,r2}	33	0.33
{r4,r5,r6}	31	0.31
{r3,r8,r9}	40	0.40
{r2,r4,r6}	63	0.63
{r3,r5,r7}	9	0.09
{r1,r4,r7}	11	0.11
{r5,r6,r7}	78	0.78
{r1.r2,r7}	89	0.89

Table 1: Performance of extraction algorithm for different rulesets

5 Evaluation and Results

5.1 Evaluation Scenario

For purposes of evaluation, we chose two different text corpora from Health and Exercise Physiology, which are not entirely similar domains, however they aren't hugely different either and thus allow an extraction algorithm to use the same set of rules for comparison. We planned to base our tests on a subset of 10 rules, $|\{r_i\}| = 10$. From the possible 2^n combinations of the rules, we carefully chose 100 that corresponded to the set $\mathbf{S}$ of rule sets. The actual value of labeled knowledge constructs in the text was $n_1 = 100$ for Health and n_2 =90 for Exercise Physiology. For each of these sets of rules, we ran the extraction algorithm configured to use these rules and extract the knowledge constructs (x_i) and compute the corresponding performance ratios, p_i. A part of these results for Health text is shown in the table 1.

5.2 Analyzing the Data

<u>Checking the Assumptions</u>:

By the very nature of the problem, we've independent sample data. This is since one particular ruleset, $\{r_i\}$, leads to a performance ratio, which is completely independent of the case when we take a negation of that particular ruleset. We also make sure that a sufficiently large sample size (100 rulesets) is undertaken for the analysis. Further, we also make sure that we've adequate cell sizes ($>$5 in most cases and non-zero in all cases).

As far as the condition of the hypothesized distribution being specified in advance is concerned, we can't say that this holds in our case since there isn't a way to know this without actually running the extracting algorithm using the respective rulesets. The condition of non-directionality becomes redundant in our case since we know how to interpret the result beforehand (rules cause the performance ratio). Further, the variables have finite values and observations are grouped in categories. Thus, most of the assumptions of the test have been met for our experimental scenario.

5.3 Calculations

We decided to demonstrate our method by calculating the significance of rule $\{r_2\}$ and the ruleset $\{r_3, r_5\}$. These rules are desribed below.

$r_2 : \{If\ only\ the\ labels\ \{nsubj, amod\}\ occur\ along\ a\ path\ P\} \longrightarrow \{extract\ P\ as\ a\ knowledge\ contruct\}$
$r_3 : \{If\ the\ labels\ \{dobj, pp\}\ occur\ consecuteively\ along\ a\ path\ P\} \longrightarrow \{extract\ P\ as\ a\ knowledge\ contruct\}$
$r_5 : \{If\ the\ label\ \{ccomp\}\ occurs\ along\ a\ path\ P\} \longrightarrow \{extract\ P\ as\ a\ knowledge\ contruct\}$

Intuitively, *nsubj* should always occurs with a *dmod* in text since it signifies a subject and object relation in linguistics. For this reason, we expect r_2 and r_3 not to affect the performance significantly. Further, *ccomp* signifies a non-trivial sentence (eg., abc claims *that* bcd is good) and thus, it's extraction requires application of finer-level of rules than r_5. Thus, our overall expectation is that these rules shoudn't affect the performance significantly when applied with other rules that are better than these in terms of extracting labelled knowledge constructs.

Using the values from the above table, we compute the following contingency tables for the two experimental texts based on the methodology explained in 4.1

Text 1:

	C_1	C_2	C_3	C_4	
$\{r_2\}$	8	5	7	9	29
$\sim\{r_2\}$	8	7	12	10	37
	16	12	19	19	66

	C_1	C_2	C_3	C_4	
$\{r_3, r_5\}$	7	8	7	6	28
$\sim\{r_3, r_5\}$	8	9	5	7	29
	15	17	12	13	57

The tables of expected counts $[n_{i.} * n_{j.}/n_{..}]$ are as shown below:

	C_1	C_2	C_3	C_4
$\{r_2\}$	7.030303	5.272727	8.348485	8.348485
$\sim\{r_2\}$	8.969697	6.727273	10.65152	10.65152

	C_1	C_2	C_3	C_4
$\{r_3, r_5\}$	7.368421	8.350877	5.894737	6.385965
$\sim\{r_3, r_5\}$	7.631579	8.649123	6.105263	6.614035

Text 2:

	C_1	C_2	C_3	C_4	
$\{r_2\}$	10	12	6	7	
$\sim\{r_2\}$	8	7	5	9	
	18	19	11	16	

	C_1	C_2	C_3	C_4	
$\{r_3, r_5\}$	11	11	5	6	33
$\sim\{r_3, r_5\}$	5	7	8	8	28
	16	18	13	14	61

Degrees of freedom for all the cases is the same, $df = (4-1)*(2-1) = 3$.

The test statistic for first case is calculated as, $\chi^2 = (8-7.03)^2/7.03 + (5-5.27)^2/5.27 + (7-8.348)^2/8.348 + (9-8.348)^2/8.348$
$+(8-8.969)^2/8.969 + (7-6.727)^2/6.727 + (12-10.6515)^2/10.6515 + (10-10.6515)^2/10.6515 = 0.7426063$.

Further, the p-value is $1 - pchisq(0.7426, 3) = 0.863$.

We now determine the p-values for all the above four cases.

For Text 1:

```
O <- data.frame(c1=c(8,8), c2=c(5,7), c3=c(7,12), c4=c(9,10));
chisq.test(O);
Pearson's Chi-squared test
X-squared = 0.743, df = 3, p-value = 0.863
```

```
O <- data.frame(c1=c(7,8), c2=c(8,9), c3=c(7,5), c4=c(6,7));
chisq.test(O);
Pearson's Chi-squared test
X-squared = 0.5184, df = 3, p-value = 0.9148
```

For Text 2:

```
O <- data.frame(c1=c(10,8), c2=c(12,7), c3=c(6,5), c4=c(7,9));
chisq.test(O);
Pearson's Chi-squared test
X-squared = 1.3281, df = 3, p-value = 0.7225
```

```
O <- data.frame(c1=c(11,5), c2=c(11,7), c3=c(5,8), c4=c(6,8));
chisq.test(O);
Pearson's Chi-squared test
X-squared = 3.7321, df = 3, p-value = 0.2919
```

From the above analysis, we clearly see that none of the p-values is significant enough to reject the null hypothesis. We interpret this observation in the following section.

6 Discussion and Future Work

Our results indicate (as per our intuition) that the presence or absence of the rule $\{r_2\}$ and the ruleset$\{r_3, r_5\}$ is independent of the performance ratio, or doesn't affect it directly. It is critical to observe that this independence is due to the presence of other rules in the ruleset used by the extraction algorithm. This gives us an indication as to which rules we may want to ignore for future analysis in analyzing text from Health and Exercise Physiology domains. Although it did not occur in this experiment, the alternate case of rejection of the null can also occur (for other rules), and in that case, we can have more interesting insights about what rules are actually affecting the performance ratio in a significant way. Based on this analysis, it would be interesting to investigate an optimal set of rules that's applicable across multiple domains (which, in this case, is two). We intend to utilize this framework in increasing the performance of generic knowledge retrieval algorithms by utilizing the insights about optimally significant rulesets.

References

[1] Alan Agresti. Introduction to categorical data analysis. *NY: John Wiley and Sons*, 1996.

[2] ed Lieberman, Bernhardt. Contemporary problems in statistics. *NY: Oxford. Section 5*, 1971.

[3] Cartic Ramakrishnan, Krys J. Kochut, and Amit P. Sheth. A framework for schema-driven relationshipdiscovery from unstructured text. *International Semantic Web Conference*, 2006.

[4] Eric W Weisstein. Chi-squared test. *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Chi-SquaredTest.html.*

[5] Zheng Zhao and Huan Liu. *Searching for Interacting Features.* 2007.