DETECTING CANCER PROGRESSION IN RADIOLOGY REPORTS

YU YAN

Abstract. The problem of detecting cancer progression in radiology text reports was analyzed based on the language of relational logic[1]. A relation extraction centric approach was proposed as the best method for solving the problem. An initial attempt at a relation extraction algorithm was presented with promising results.

1. Introduction

1.1. **Motivations**

. The benefits of automatically detecting cancer progression in radiology reports are obvious. I worked with on this project with Dr. Rubin in the radiology department.

1.2. **Structure Of Data**

. A *token* is the most basic unit in this report's world. The string form that we find it in is just one of its many attributes. I trust that you know what I am talking about and will not delve into the pedantic question of what it really is. I will however define all other concepts in terms of the token.

Definition 1. Text

A *text* is a collection of tokens

Definition 2. Attributes

An *attribute* is a function whose domain is the set of all tokens

Definition 3. Relation[1]

A *n-ary relation* is a function with arity n and whose domain is an n-tuple of the set of all tokens, attributes and object constants

Definition 4. Rule

A *rule* is a logical statement involving relations

Example 5. Examples

(1) Novels, reviews and radiology reports are examples of texts

(2) The string value, the position value and the part-of-speech value of a token are examples of attributes
(3) A triple of tokens lying in the same sentence is an example of a ternary relation

1.3. **Problem Definition**

.

Definition 6. Cancer Progression

Let t_1, t_2 be 2 time points in chronological order and let p be a patient,
The cancer in p has progressed from t_1 to t_2 if any of the following is true:

(1) There are more cancerous nodules in p at t_2 than t_1
(2) The cancerous nodules in p at t_2 have increased in size since t_1

Corollary 7. *Two ways in which radiology reports can be used*

Let R be a collection of radiology reports for a patient p, each written at different times,

R indicates a cancer progression for p if any of the following is true:

(1) *There exists $r_1, r_2 \in R$ where r_2 was written later than r_1 such that any of the following is true(to make things even less intuitive they are not mutually exclusive):*
   (a) *There are more cancerous nodules described in r_2 than r_1*
   (b) *There exists a cancer nodule described in r_2 having a size bigger than the same nodule described in r_1*
(2) *There exists $x \in R$ such that there are explicit comments written in x suggesting a disease progression.*

Note the following differences between options 1 and 2 above:

(1) For option 2, the radiologist has manually done the comparison and is providing textual comments or hints of progression, whereas in option 1 these comments may not exist.
(2) If option 2 is true, only that one radiology report is needed to indicate disease progression. Rather than having to look through and compare many reports, only 1 report is needed.

This project will only deal with cases where option 2 is viable. This means that a significant amount of data preprocessing has to happen to filter out reports that do not satisfy option 2 above.

Example 8. Option 2 Report

The following are examples of sentences in an option 2 report:

(1) "There is a disease progression..."
(2) "The nodule has enlarged since..."
(3) "New nodules are observed since..."

At this point you should be rolling your eyes and making the following comments:

(1) Why don't they(the vagueness of this pronoun is intentional) just figure out a structured way for the radiologist to indicate the presence of progression?
   (a) I don't know.
(2) Are you going to ignore the much harder task of tracking the status of the various nodules across different reports by information extraction and instead just try to figure out if a progression exists based on a radiologist's comments/opinions?
   (a) Yes
(3) Isn't it too easy?

1.4. **The GRL model**

. Below is a modeling of the problem using concepts from general relational logic(GRL).

Let x, y, z be tokens in a text T, *lemma* be an attribute function mapping a token to its lemma value, D be the dependency parse tree of T, consider the following relations:

Dependency Relations:

(1) $depends\,(x, y)$ if x is a child node of y in D
(2) $dependent\,(x, y)$ if any of the following is true:
   (a) $depends\,(x, y)$
   (b) $\exists z \in T$ such that $depends\,(x, z)$ and $dependent\,(z, y)$
(3) $related\,(x, y)$ if any of the following is true:
   (a) $dependent\,(x, y)$
   (b) $dependent\,(y, x)$

Semantic Relations:

(1) $corefer\,(x, y)$ if x and y corefer
(2) $synonym\,(x, y)$ if $lemma\,(x)$ and $lemma\,(y)$ are synonyms
(3) $negated\,(x)$ if $\exists z \in T$ such that all of the following are true:
   (a) $not\,(z)$
   (b) $dependent\,(z, x)$
(4) $equal\,(x, y)$ if any of the following is true:
   (a) All of the following are true:
      (i) $synonym\,(x, y)$
      (ii) $\neg negation\,(x)$
   (b) $\exists w \in T$ such that all of the following are true:
      (i) $corefer\,(w, x)$
      (ii) $equal\,(w, y)$

Lemma Relations:

(1) $not\,(x)$ if $synonym\,(x, z)$ for $lemma\,(z) =$ "no"
(2) $new\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "new"
(3) $nodule\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "nodule"
(4) $disease\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "disease"

(5) $grow\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "grow"
(6) $increase\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "increase"
(7) $progress\,(x)$ if $equals\,(x, z)$ for $lemma\,(z) =$ "progress"

Disease Relations:

(1) $newnodules\,(x)$ if all of the following are true:
    (a) $nodule\,(x)$
    (b) $\exists y \in T$ such that all of the following are true:
        (i) $new\,(y)$
        (ii) $related\,(x, y)$
(2) $nodulegrowth\,(x)$ if all of the following are true:
    (a) $nodule\,(x)$
    (b) $\exists y \in T$ such that all of the following are true:
        (i) $grow\,(y)$
        (ii) $related\,(x, y)$
(3) $diseaseprogression\,(x)$ if all of the following are true:
    (a) $nodule\,(x)$
    (b) $\exists y \in T$ such that all of the following are true:
        (i) $progress\,(y)$
        (ii) $related\,(x, y)$
(4) $moresizeornumber\,(x)$ if all of the following are true:
    (a) $nodule\,(x)$
    (b) $\exists y \in T$ such that all of the following are true:
        (i) $increase\,(y)$
        (ii) $related\,(x, y)$

Final Relation:

(1) $progression\,(T)$ if any of the following is true:
    (a) $\exists x \in T$ such that $newnodules\,(x)$
    (b) $\exists x \in T$ such that $nodulegrowth\,(x)$
    (c) $\exists x \in T$ such that $diseaseprogression\,(x)$
    (d) $\exists x \in T$ such that $moresizeornumber\,(x)$

Given a report T, our task is to determine the truth value of $progression\,(T)$

1.5. **Basic assumption of the project**

. The GRL model is by no means a complete description of the problem. For example, the *progression* relation has only 4 rules, and it seems that they are insufficient for modeling the whole problem. As a result, if we apply the GRL model to our problem, we would expect to end up with a high precision and low recall. This is not to mention the trouble one has to go through to code the GRL model precisely. The only point of introducing the incomplete GRL model is to inspire the following proposition, which will result in the final approach taken in this project.

Proposition 9. *Unary and Binary rules are enough*

Let R be an option-2 radiology report, there is a set S of rules(composed of binary or unary relations on the tokens in R) such that S is logically equivalent to disease progression in R

Remark 10. Intuitions(not proof) of the above proposition

If you trace your way through the GRL model starting with the *progression* relation, you would notice that you are only encountering binary and unary relations. If the GRL model is perfect, then the proposition above is automatically correct. However, we suspect that the GRL model is not perfect. In particular, we suspect that more relations are needed. The central assumption I make in this problem is this: *whatever new relations are needed to perfect the GRL model, they must still be unary or binary relations.* The preceding phrase in italics is actually equivalent to the proposition. Of course, a rule based classifier based on a perfected GRL model will achieve a 100% accuracy on this problem.

From now on, I will call the set of rules required for the proposition above the "set of GRL rules" and the associated set of relations the "set of GRL relations"

1.6. **Application of Naive Bayes Classifier**

. Let K be a superset of the set of GRL relations. Let F, G be sets of indicator features associated with K and the GRL relations respectively. The Naive Bayes method[2] makes some assumptions which result in the following equation:

$c = argmax\left(P\left(c\right)\Pi_{f\in F}P\left(f \mid c\right)\right)$ where $c = 1$ if there is a cancer progression and 0 otherwise.

Consider $f_i \in F$, if f_i is not associated with a GRL relation, when we train the classifer on a large dataset, $P\left(f_i \mid c = 1\right) \simeq P\left(f_i \mid c = 0\right)$ because the relation should appear with approximately the same frequency regardless of the class(if this is not true, it means that it is not true that f_i is not associated with a GRL relation, contradicting the premise). Hence, we get:

$c = argmax\left(P\left(c\right)\Pi_{f\in G}P\left(f \mid c\right)\right)$ meaning that the classifier depends only on the GRL features. This results in the first part of the proposition below.

On the other hand, since by construction some GRL features will never indicate 1 for a non-progression case, $argmax\left(P\left(c\right)\Pi_{f\in G}P\left(f \mid c\right)\right)$ should closely approximate the perfect rule-based classifier when all the GRL rules are found. This forms the second part of the proposition below.

Proposition 11. *More is not worse*

Let K be a superset of the set of GRL relations. Let R be the set of all possible radiology reports and let F, G be sets of indicator features associated with K and the GRL relations respectively,

whenever they are trained on R,

(1) *a Naive Bayes classifier with feature set F is equivalent to a Naive Bayes classifier with feature set G*

(2) *a Naive Bayes classifier with feature set F is equivalent to the rule based classifier based on the perfect GRL model.*

The proposition calls for us to develop an algorithm that will return a set containing all the GRL relations, even if the set is large. Of course, the set of possible relations of maximum arity 2 is infinite, so simply "including everything" does not work.

Although the proposition is made specifically with respect to the Naive Bayes classifier, we will tolerate the intuition that a simlar argument holds for other types of classifiers and look at the results.

1.7. **Summary of discussion**

. In short, when we make the following assumptions:

(1) A perfect rule based classifier based on a perfect GRL classifier can achieve 100% accuracy on this problem
(2) The perfect GRL model contains relations with arity at most 2

We arrive at the following conclusion:

(1) The relation extraction algorithm which extracts all the GRL relations is central to the problem
(2) Trained on a large enough dataset and using a feature set containing the GRL relations, the Naive Bayes classifier can achieve a perfect accuracy just like the perfect GRL rule based classifier

1.8. **A relation extraction algorithm**

. I have come up with several relation extraction algorithm for the problem but will only present here the one which achieved the best result.

(1) Do dependency parsing on the entire collection of radiology reports
(2) For each dependency pair, if it contains a word that is synonymous with any of the following set of words (nodule,new,increase,disease,progress), save the relation, disregarding the order in which the words appear
(3) Remove all relations that appeared only once

There are various problems associated with this algorithm, which means that it cannot possibly extract all the useful GRL relations. These problems include:

(1) Absence of unary relations, coreference resolution and negation handling
(2) No interactions between relations. The GRL model presents a hierarchy of relations where some relations depend on others. This algorithm disregards the hierarchy completely and stays at the dependency pair level

For all its shortcomings, the algorithm has produced impressive results, especially compared with alternatives. This suggests that the relation extraction centric approach that I suggest in this paper is indeed appropriate for the problem. The results will be presented later.

2. Implementation Details

2.1. **Programming language and libraries**

.

(1) The relation extraction code was written in java with a heavy use of the following libraries:
   (a) Stanford CoreNLP
   (b) OpenNLP
(2) The data preprocessing and machine learning code was written in python with a heavy use of the following libraries:
   (a) NLTK
   (b) Scikit-Learn

2.2. **Data**

. The data comes from radiology reports at the Stanford Hospital. There are about 2000 reports available, 900 left after preprocessing, 158 labeled data, about 40% of which contains progression(the actual data before preprocessing is much more skewed towards no progression). Each report comes with metadata about its modality, body part as well as date.

2.3. **Data Preprocessing**

. Since the project deals only with single radiology reports, the non-option-2 reports(as discussed in the introduction section) have to be discarded. To acheve this, significant amount of data preprocessing took place, with an algorithm outlined below:

For each report:

(1) Section segmentation to isolate the part of the report containing information about the report that the radiologist is currently comparing to. We call this the comparison section.
(2) Named entity recognition as well as extraction of modality, date as well as body part information within the comparison section. The task was accomplished by using regular expressions.
(3) Comparison of the extracted entities with the metadata and discard report if:
   (a) The modality of the current report and compared report do not match
   (b) The body part discussed of the 2 reports do not match
   (c) The date of the 2 reports are within a month of each other
(4) Section segmentation to isolate the part of the report where the radiologist directly compares the current report with a previous report. We call this the impression section.

The resulting option-2 reports represent less than 50% of the total number of reports in this dataset.

2.4. **Relation Extraction**

. The relation extraction algorithm was already described in the introduction section. Here I will provide more details about the code:

(1) Dependency parsing and word stemming was done by the Stanford CoreNLP package.
(2) Word lemmatization was done by the OpenNLP lemmatizer
(3) At the time this report was written, I have yet to be approved to use MetaMap, which is a java medical thesaurus that I can use to identify the synonym relation. The results presented here are expected to improve once that is used.

2.5. **Machine Learning**

. The dataset was first separated as follows:

(1) 80% of the shuffled data constitutes our training set
(2) The remaining data becomes our test set. This set is used only once to report the final test result of the best classifier.

On the training set, 10 fold cross validation was used and the best model was picked based on the f1 score. In order to prevent training on the test set, the test results for all other classifiers are not even evaluated. No hyperparameter tuning was done due to time constraints. Results using both relation features and single word features(using every word from corpora) are presented.

3. Results

3.1. **Model Selection**

. The classifiers used can all be found in the scikit-learn package

Classifier - Using relation features	F1 Score
MultinomialNB	0.73
ExtraTreesClassifier	0.55
LogisticRegressionCV	0.67
svm.SVC	0.0

Classifier - Using single word features(bag of words)	F1 Score
MultinomialNB	0.44
ExtraTreesClassifier	0.39
LogisticRegressionCV	0.32
svm.SVC	0.0

3.2. **Best model test results**

.

Best Classifier	Accuracy	F1	Precision	Recall
MultinomialNB	0.81	0.73	0.73	0.82

3.3. **Discussion of results**

. The results show decisively that the relation features are better than single word features. The Naive Bayes Classifier turned out to be much better than the other discriminative classifiers. This may reflect the fact that the derivations used in this project were based on the Naive Bayes Classifier and

its results may not extend to other classifiers. The SVM classifier had a 0 F1 score because it declared all the reports as negative. The following factors placed could contribute to lower scores:

(1) Lack of hyperparameter tuning. All the default scikit-learn values were used due to time constraints
(2) Small dataset size. One central premise in this approach is that the dataset size has to be large. With only 158 labeled data, the classifiers cannot be expected to approach the GRL rule based classifier in performance
(3) Lower quality data labeling. As a non-radiologist, my labeling of the data can be expected to contain errors.

In light of these limiting factors, the 81% test accuracy is excusable, although a 100% is the ultimate goal.

4. Conclusion and Future Work

The central thesis of the paper that a relation extraction centric approach can ultimately solve the problem remains unrefuted based on the tentative results. It seems to suggest that any future work should focus on bettering relation extraction. I am also interested in seeing how far the hypothesis extends beyond this one application task. Finally, I am interested in further understanding this and other related problems in terms of computational semantics.

References

[1] "Introduction to Logic." Introduction to Logic. Web. 08 Dec. 2015. <http://logic.stanford.edu/intrologic/notes/notes.html>.

[2] Wikipedia. Wikimedia Foundation, Web. 08 Dec. 2015. <https://en.wikipedia.org/wiki/Naive_Bayes_classifier>