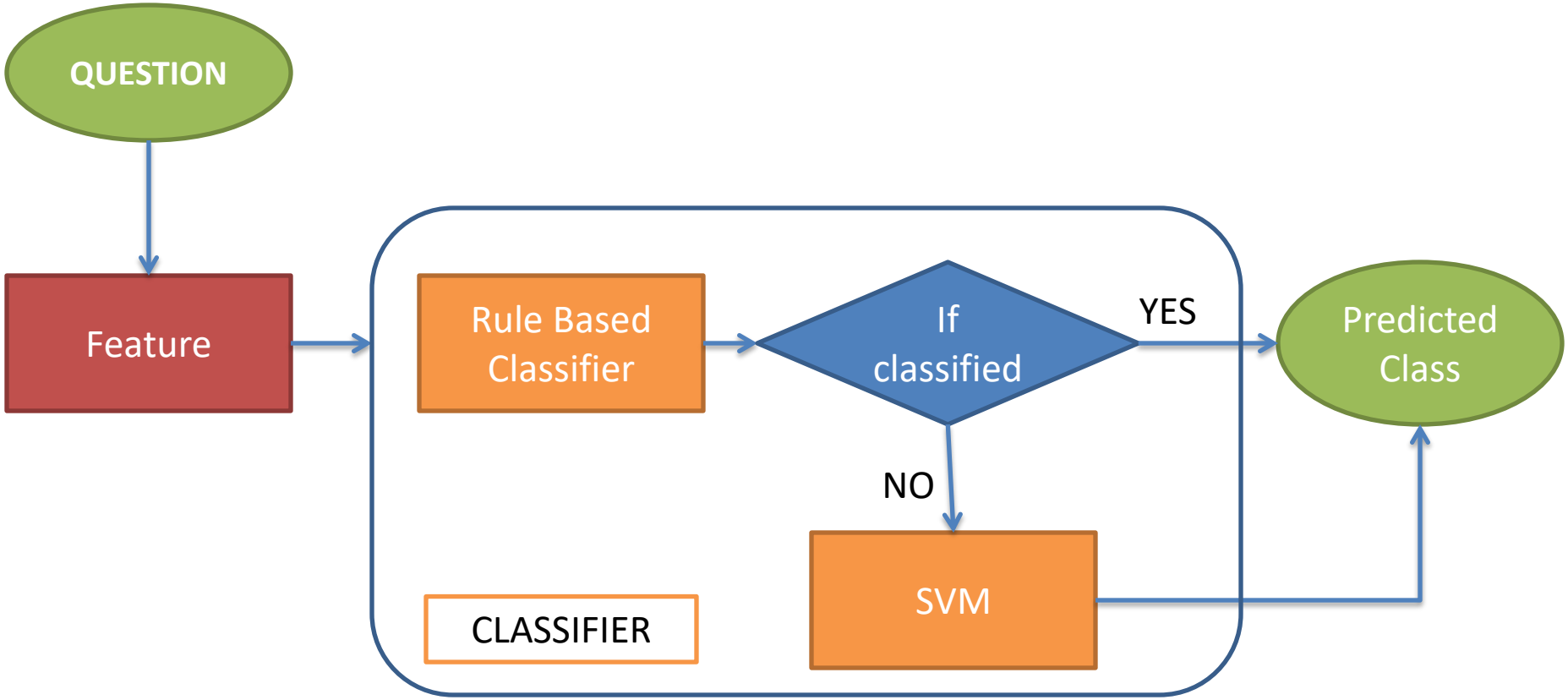
Amazon-Qlearn

https://inclass.kaggle.com/c/amazon-qlearn/

Aashraya Sachdeva
https://www.kaggle.com/aashsach
aashrayasachdeva2@gmail.com
+91-9582254549

Software Engineering And Analysis Lab
Department of CSA
Indian Institute of Science
Bangalore, 560012



Architecture

Feature Extraction

The input question text is represented by following vectors:

1. Input text/ Text tokens (**TOKS**): Its just input text tokenized to words.

2. Part-of-Speech (**POS**) of text: Part of speech tagging for every token[1].

3. Entity Tagging(**EnT**): Assigns labels to contiguous spans of tokens[2].

4. Noun-vector (**NV**): All words in **EnT** are added. Every noun word that's not in **EnT** is mapped against a ***synonym dictionar***y. If matched, the synonym is kept in Noun-vector else dropped. All words which are not Noun are dropped. Finally, all the words in **NV** are stemmed.

Synonym Dictionary

A dictionary is created which maps common words to their closest synonyms using one of following ways:

1. Xin and Roth[3] has a list of synonyms available at [4].

2. A synset of phrases : "person", "place", "substance", "quantity" using wordnet[5] was created. A Noun is classified into to closest synset group (if not found in above list) if its distance is less than 6.

[1] https://spacy.io/docs/usage/pos-tagging
[2] https://spacy.io/docs/usage/entity-recognition
[3] Xin Li, Dan Roth, Learning Question Classifiers. COLING'02, Aug., 2002.
[4] http://cogcomp.cs.illinois.edu/Data/QA/QC/QC.tar
[5] https://wordnet.princeton.edu/

Rule Based Classifier

Rules are created manually using following:
1. Unigram and bigram frequency distribution of TOKS and POS in Test Data.
2. English Domain knowledge.

Class 0: Abbreviation
1. An acronym and "***mean***" in TOKS
2. An acronym and "**What is**" in TOKS
3. TOKS contains any of : "**stand for**" ,"**abbreviation**", **acronym**".

Class 1: Human
1. TOKS starts with any of: "**Whom**", "**Whose**"," **Who**".

Class 2: Location
1. TOKS starts with "**Where**".
2. TOKS starts with "**Is there**" and TOKS contains "**place**".
3. TOKS starts with "**Which**" and TOKS contains "**place**".

Class 3: Description
1. TOKS starts with any of: "**Why**", "**If**"., "**Define**", "**Describe**", "**Explain**", "**Suggest", "Give reason".**
2. TOKS starts with "**Which**" and not followed by a POS-**JJ** or POS-**RB**.

Class 4: Entity:
NO RULES

Class 5: Description
1. TOKS starts with "**Which**" and followed by a POS-**JJ** or POS-**RB**.

Support Vector Machine (SVM)

Class Prediction:

1. SVM is trained for classes 1,2,3,4,5. Class 0 data is excluded from training.
2. Thus does not predict class 0.
3. Class 0 prediction is done only by Rule based Classifier.

Classifier: ONE-VS-REST

Input Feature: Bag-of-word representation of NV and TOK

Kernel: Linear

Regularization Constant (C): 1.0.

Training Procedure:

1. Cross-validation Grid search: C was estimated by training SVM on 80% of training data and validation on rest 20%.
2. Train SVM on entire train Data with estimated C.