Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status

- ***Richa Kushwaha*** *(20172056)*
- ***Sudheer Achary*** *(20161076)*
- ***Swati Tyagi*** *(20172096)*
- ***P. Sai Vasishth*** *(201501179)*

Introduction

The paper proposes a strategy for the summarization of scientific articles using the rhetorical status of statements in an article. It also present an algorithm that selects content from unseen articles and classifies it into a fixed set of seven rhetorical categories, on the basis of the annotated training material.

Aim Of The Project

- This project builds towards automatic summarisation of scientific papers. To that end, we aim to classify each sentence within the research paper as one of the rhetorical categories thus aiming to disseminate information that can be easily consumed by the public.
- The system can thus serve as great inroads into better scientific understanding among researchers, academicians and the general population. We divided the tasks :
    - ***Compile a dataset*** ~50 papers with rhetorical information tagged as described in the paper.
    - ***Implement an algorithm*** - That, on the basis of the annotated training material, selects content from unseen articles and classifies it into fixed set of seven rhetorical categories. This algorithm is described in the paper.
- The output of this extraction and classification system can be viewed as a single-document summary in its own right. Alternatively, it provides starting material for the generation of task-oriented and user-tailored summaries designed to give users an overview of a scientific field.
- We have come up with an idea of a deep learning model to solve the problem.

Classes and Annotations

As the first task we need to annotate each sentences in the 50 research papers. So we need to define rhetorical status so as to describe the content in scientific articles. It should both capture generalizations about the nature of scientific texts and also provide the right kind of information to enable the construction of better summaries for a practical application. The rhetorical status hence captures following details:

- ***Problem structure***: Research is often described as a problem-solving activity. *It is defined in the project using some annotations like [AIM].*
- ***Intellectual attribution:*** It consists of what the new contribution is, as opposed to previous work and background material (generally accepted statements). *It's used in annotations like [CONTRAST], [BASIS], [BACKGROUND].*
- ***Scientific argumentation:*** In contrast to the view of science as a disinterested, researchers like Swales (1990) have long claimed that there is a strong social aspect to science, because the success of a researcher is correlated with her ability to convince the field of the quality of her work and the validity of her arguments.
- ***Attitude toward other people's work:*** It consists of how authors include reference to other work into their argument. It is portrayed as a rival approach, or as an approach contributing parts of the authors' own solution. *It can be used using [CONTRAST], [BASIS].*
- ***Citations and Relatedness:*** We concentrate on two citation contexts that are particularly important for the information needs of researchers:
    - Contexts in which an article is cited negatively or contrastively using *[CONTRAST]*.
    - Contexts in which an article is cited positively or in which the authors state that their own work originates from the cited work using *[BASIS]*.

Rhetorical Annotation Scheme

The rhetorical annotation scheme (cf. Table 1) encodes the aspects of scientific argumentation, and relatedness to other work and all other basis described in the previous section. The following annotations are used :

Table 1
Annotation scheme for rhetorical status.

AIM	Specific research goal of the current paper
TEXTUAL	Statements about section structure
OWN	(Neutral) description of own work presented in current paper: Methodology, results, discussion
BACKGROUND	Generally accepted scientific background
CONTRAST	Statements of comparison with or contrast to other work; weaknesses of other work
BASIS	Statements of agreement with other work or continuation of other work
OTHER	(Neutral) description of other researchers' work

Fig 1. Table 1-Annotation Scheme for rhetorical status

As a corpora we have 80 annotated research papers collected from the CMP_LG archive (CMP_LG 1994) from which we have used 50 for training and 30 for testing. The following Decision Tree can be used for rhetorical annotation :

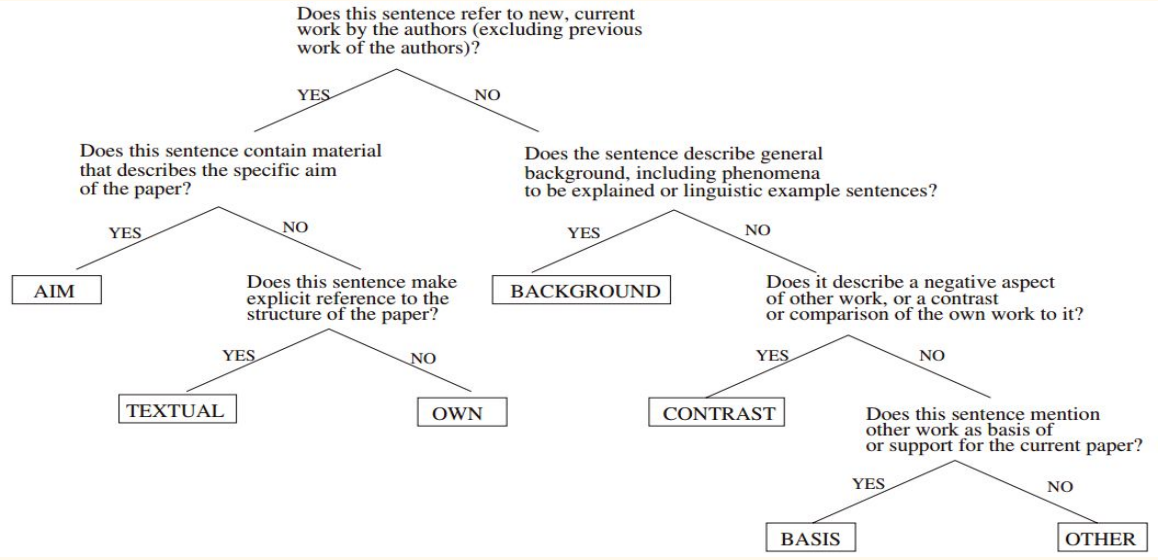


Fig 2. Decision Tree for rhetorical annotation

Features and The Model

Few sentential features were collected for each sentence. Learning is supervised:

- In the training phase, associations between these features and the seven target categories are learned.
- In the testing phase, the trained model provides the probability of each target category for each sentence of unseen text, on the basis of the sentential features identified for the sentence.

The features are borrowed from the text extraction literature (Paice 1990) or related tasks and adapted to the problem of determining rhetorical status. Overview and description of the collected features and feature types is explained as follows:

- ***Absolute location of a sentence***
    - ***Loc*** Gives position of sentence in relation to 10 segments .i.e A-J
        - The observation was that the size of rhetorical zones depends on location, with smaller rhetorical zones occurring toward the beginning and the end of the article. The location values are assigned in the following fashion:
        - The article is divided into 20 equal parts, counting sentences.
        - Sentences occurring in parts 1, 2, 3, 4, 19, and 20 receive the values A, B, C, D, I, and J, respectively.
        - Parts 5 and 6 are pooled, and sentences occurring in them are given the value E; the same procedure is applied to parts 15 and 16 (value G) and 17 and 18 (value H).
        - The remaining sentences in the middle (parts 7–14) all receive the value F.
    - ***Section Struct*** Relative and absolute position of sentence within section (e.g., first sentence in section or somewhere in second third). It can have 7 values.
        - The section location feature divides each section into three parts and
        - assigns seven values: first sentence, last sentence, second or third sentence, second-last or third-last sentence, or else either somewhere in the first, second, or last third of the section.

- ***Para Struct*** Relative position of sentence within a paragraph. The values will be: Initial, Medial, Final. Sentences are distinguished into those leading or ending a paragraph and all others.
- ***Headline*** Type of headline of current section.
    - Whenever a prototypical headline is recognized (using a set of regular expressions), it is classified into one of the following 15 classes: Introduction, Implementation, Example, Conclusion, Result, Evaluation, Solution, Experiment, Discussion, Method, Problems, Related Work, Data, Further Work, Problem Statement.
    - If none of the patterns match, the value Non-Prototypical is assigned.

- ***Sentence Length Content features***
    - ***Length*** indicates if the sentence longer than a certain threshold(12 in our case), measured in words and hence, the value can be either Yes or No.
    - ***Title*** indicates if the sentence contain words also occurring in the title or headlines. Value can be : Yes or No.
    - ***Tf-Idf*** Tells if the sentence contain "significant terms" as determined by the *Tf-Idf* measure. Values can be : Yes or No.
        - The *Tf-Idf* formula assigns high values to words that occur frequently in one document, but rarely in the overall collection of documents.
        - We used the 18 highest-scoring *Tf-Idf* words and classify sentences into those that contain one or more of these words and those that do not.
- ***Citations***
    - ***Cit*** indicates if the sentence contain a citation or the name of an author contained in the reference list. If it contains a citation, is it a self-citation? Whereabouts in the sentence does the citation occur? The value can be {Citation (self), Citation (other), Author Name, or None} × {Beginning, Middle, End}
        - We used a recognizer for formal citations.
        - It parses the reference list at the end of the article and determines whether a citation is a self-citation (i.e., if there is an overlap between the names of the cited researchers and the authors of the current paper), and it also finds occurrences of authors' names in running text, but outside of formal citation contexts (e.g., Chomsky also claims that . . .).
        - The citation feature reports whether a sentence contains an author name, a citation, or nothing. If it contains a citation, the value records whether it is a self-citation and also records the location of the citation in the sentence (in the beginning, the middle, or the end).

- ***History***
    - ***History*** indicates most probable previous category that can be 7 Target Categories + “BEGIN”.

Annotation and Observation

We have also annotated 15 papers which were given to us and were based on psychological diseases. Some points to note about the data and some issues with the annotation process and which refers to the problem statement:

- **A minimum knowledge about the domain related to the paper is must.** This will help to know to differentiate the BACKGROUND labels specially. Also, it may lead to confusion in understanding if a sentence refers to author’s work or new work also (at worst cases) if the annotator doesn’t have domain knowledge.
- **A basic understanding about the paper(at least abstract) is must.** Annotating the paper is equivalent to reading it and understanding the differences between new work and old work, and understanding the intricacies in the flow of ideas.
- **Future context also matters.** Sometimes, author just mentions some methods are present in some sentence. He doesn’t really support them, neither oppose them (OTHER is preferable). But, he does so in next sentence or after a few. So, if he opposes it, it is better to put it in CONTRAST. So, telling something solely by a sentence is not right. Previous and Future contexts also matter.
- **Summary Papers.** Some papers are just papers meant for summarising papers in some research topic . Such papers are highly confusing to annotate. The work done by author here is collecting data and analysing. The analysis he mentions maybe his own, or the paper he mentions. Here the definition of work done by author is different. The work he did is collecting data and explaining it. The explanation provided by him is mostly not know if it is his own or not, or it varies with how we define “work done by author”. The sentence can be his **own**(if I say he is explaining, or thus summarising the paper), or can be a **basis**(if I say he is taking the help of a paper to convey something).

Implementation and Results

We use ***xml*** python package to parse through annotated data which is presented in *xml* format. Using this we almost extract every feature presented in paper. The following feature extractors are implements.

Features

- ***Absolute Location*** It describes the positional information of a sentence among equally partitioned 20 segments.
  Following Tags are given = {A, B, … J} accordingly
- ***Explicit Structure*** It describes the structural information of a sentence within a section.
  Following Tags are given = {FIRST, SECOND, … SOMEWHERE} accordingly
- ***Paragraph Structure*** It describes the positional information of a sentence within a paragraph.
  Following Tags are given = {INITIAL, MEDIAL, FINAL} accordingly
- ***Title*** It describes if any content of title present in a sentence.
  Following Tags are given = {YES, NO} accordingly
- ***Length*** As the name suggests it specifies if sentence is more than a specified amount of length (measured w.r.t words)
  Following Tags are given = {YES, NO} accordingly
- ***Citation & Reference*** It describes the presence of a citation in the sentences, if yes then it describes does it refer to the author itself or any other author. It also describes the positional information of citation within a sentence.
  Following Tags are given = { {YES, NO} X {YES, NO} X {SELF, OTHER} } accordingly
- ***Headline*** It describes if the section is among one of the 15 prototypical sections as such *Introduction, Implementation … etc*. mentioned in the paper
  Following Tags are given = {YES, NO} accordingly
- ***History*** It describes the category/class that was assigned with it, which carries a kind of temporal information as they occur sequentially
  Following Tags are given = {YES, NO} accordingly
- ***Tf-Idf*** It describes the importance of a sentence, by describing the presence of top n - *Idf* scored words in sentence chosen from vocabulary
  Following Tags are given = {YES, NO} accordingly

Classifiers

$$P(C \mid F_0, \ldots, F_{n-1}) \approx P(C) \frac{\prod_{j=0}^{n-1} P(F_j \mid C)}{\prod_{j=0}^{n-1} P(F_j)}$$

$P(C \mid F_0, \ldots, F_{n-1})$:	Probability that a sentence has target category C, given its feature values $F_0, \ldots, F_{n-1}$;
$P(C)$:	(Overall) probability of category C;
$P(F_j \mid C)$:	Probability of feature-value pair F_j, given that the sentence is of target category C;
$P(F_j)$:	Probability of feature value F_j;

Fig. 3 Naive Bayesian Classifier

- ***Naive Bayes*** We used naive bayes classifier with different distribution assumption on likelihood ratio. So, preprocessed data of each sentence that represents 14 feature vectors gets feeded into one of these models and predicts the class assigned to it. We had assumed 4 distributions ***Bernoulli, Multinomial, Gaussian, Compliment***. Among which *Bernoulli* gives a highest accuracy of ~80% where as *Gaussian* gives a least accuracy of ~25%. The each of these model was trained and tested on different sets of data, such that ratio of train test split was maintained to ~0.8.
- ***Deep Network*** We have also tried a deep learning based sentence classification, Where we feed word2vec features rather than those mentioned in the paper. It provides an accuracy ~71%. But drawbacks of this *Deep Network* are explained in latter sections.

Results

- Some of the plots show confusion matrix that describes true positive/negatives distribution over false positive/negatives, while others show histograms that describes the distribution of the predicted classes

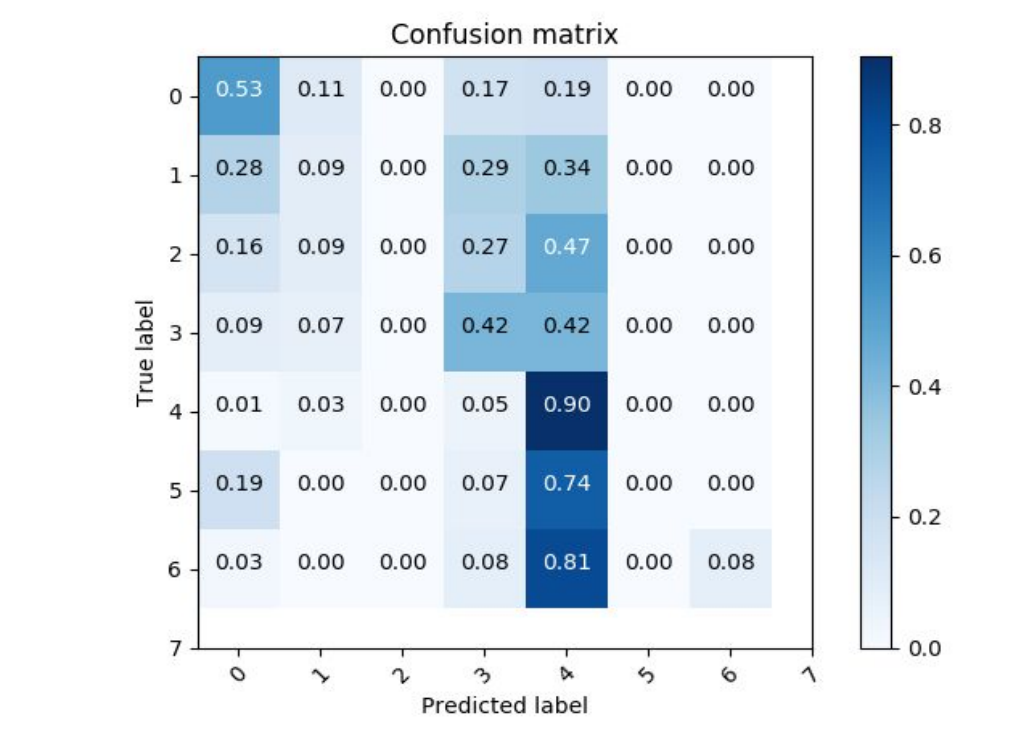


Fig. 4 ***Multinomial Naive Bayes*** This model predicts with an accuracy of *~72%*

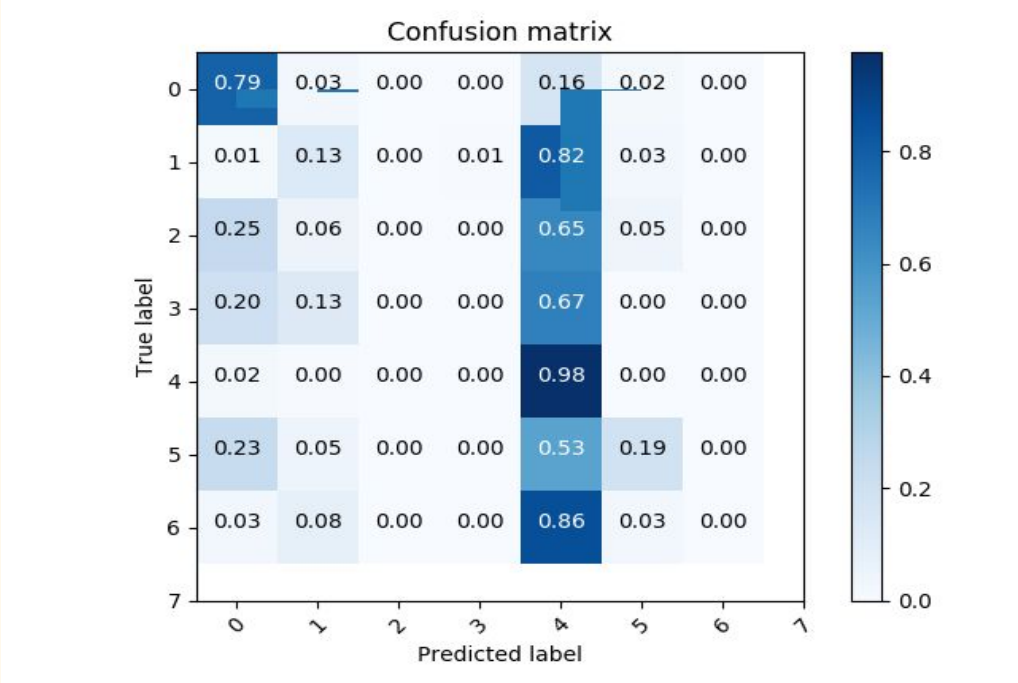


Fig 5. ***Bernoulli Naive Bayes*** This model predicts with an accuracy of *~80%*

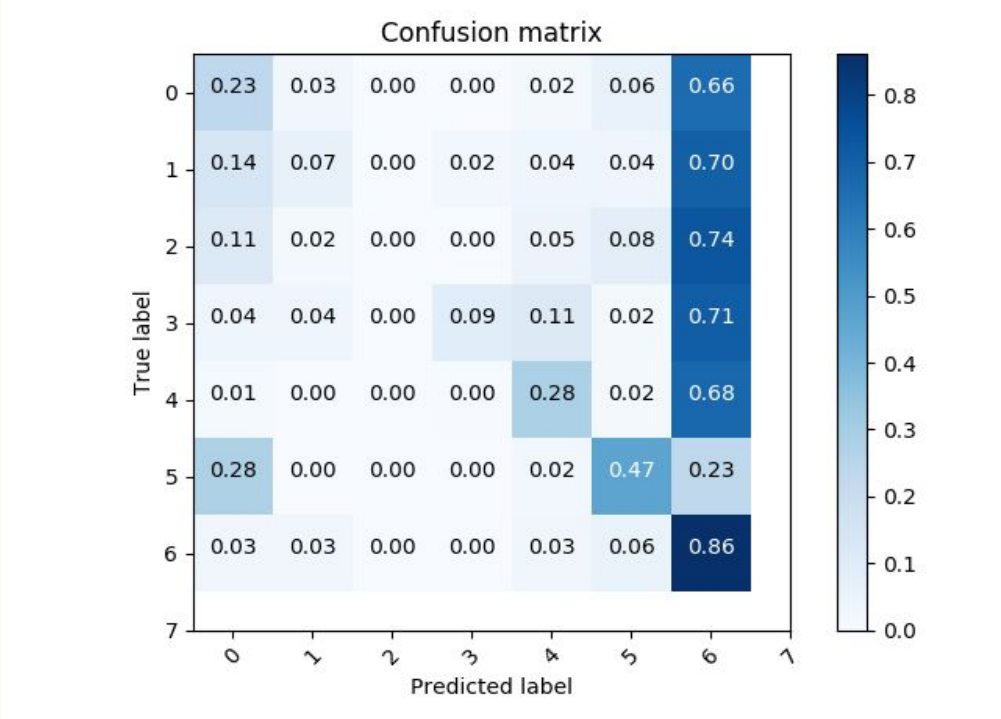


Fig 6. ***Gaussian Naive Bayes*** This model predicts with an accuracy of *~25%*

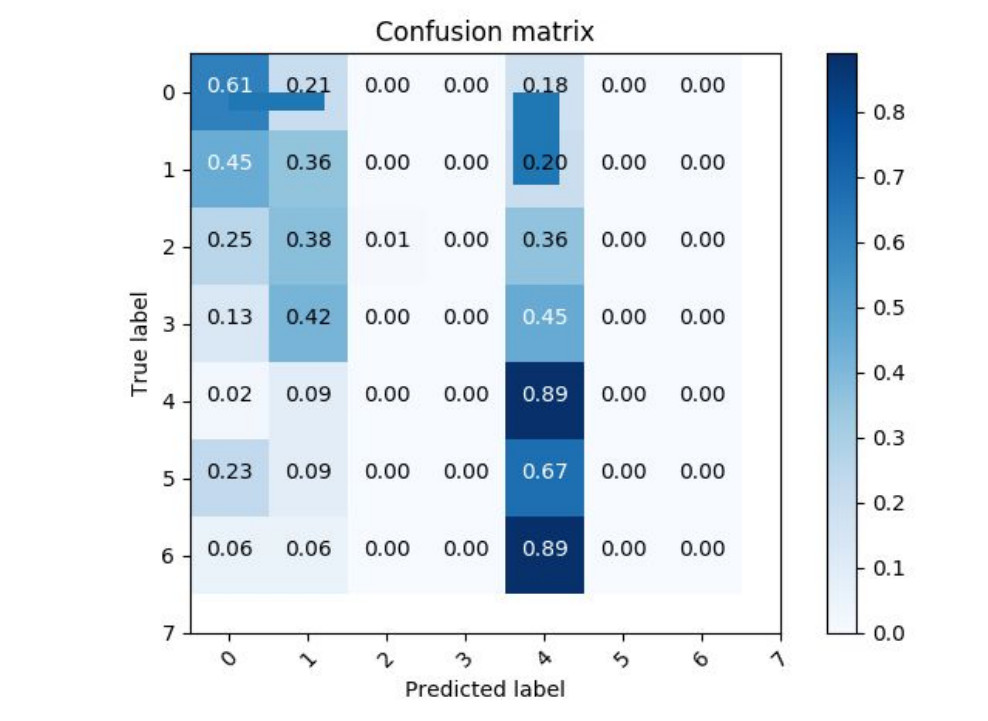


Fig.7 ***Complement Naive Bayes*** This model predicts with an accuracy of *~73%*

Deep Learning

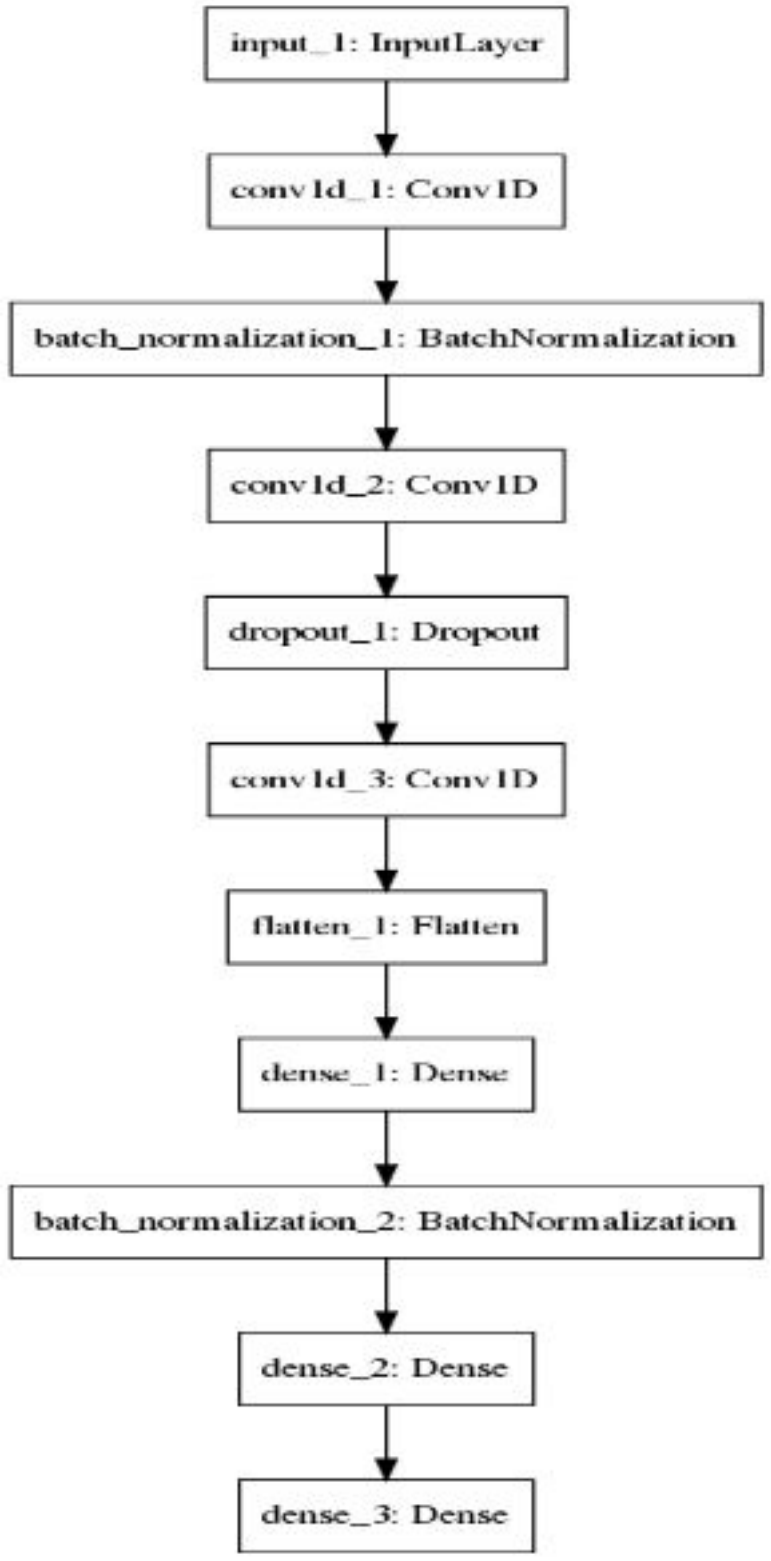
As suggested in paper each sentence is extracted with 14 significant features, namely absolute location, paragraph structure etc. We suggest a different feature representation for each sentence using ***word embeddings***. We used ***stanford GloVe embeddings*** to represent a sequential feature vector for each sentence.

We used 50 dimensional *GloVe embedding* feature vector for each word in sentence and concatenate them along to make a large feature vector for each sentence. As the sentences vary with length these feature vectors needs to be padded/truncated as ones choice to make feature vectors trainable for *Deep Network*.

After preprocessing the text to features, we then feed it to a *Deep Network.* We used a *Convolutional* layers with different set of activations & filters, followed by *Dropouts* & *Batch Normalization* finally stacked with *Dense* layers for one hot categorical classification.

We use *Adamax* optimizer for gradient descent, *categorical cross entropy* loss as it is an multiclass classification problem. The main problem we faced for less data which makes the model to learn the most probable class rather than classify based on sentencial features. As *OWN* labels are densely populated and there is an great imbalance among the classes it's a harder problem to do this with less data. It gets an *accuracy* of ~71% even with most probable class prediction.

Architecture



Future Work and Improvements

- As the occurence of each word has influence on other as it is a sentence instead of using *Deep Network* as mentioned above we propose the idea of using temporal information. Here we pad each word vector of a sentence to some assumed dimension and feed it to *Recurrent Network* like ***LSTM*** *(Long Short Term Memory),* ***GRU*** *(Gated Recurrent Networks)* networks which use temporal information also to classify a sentence to one of its rhetorical categories. Even these networks might have the issue with imbalanced data distribution among classes of sentences, along with it these networks require lot of training data which we don't have either.
- As of now we have only tried this on 50 dimensional *GloVe embeddings,* This might get improved with a large dimension such as 100, 300 … etc which makes classifiers harder to train.