*International Journal on Emerging Technologies* (Special Issue on ICRIET-2016) **7**(2): 199-206(2016)

**ISSN No. (Print) : 0975-8364**
**ISSN No. (Online) : 2249-3255**

# Heuristic Sentence Boundary Detection and Classification

***C. Gnana Chithra* and Dr. E. Ramaraj*****
**Equity Research Consultant, Angeeras Securities, Chennai.*
***Professor, Department of Computer science and Engineering,*
*Alagappa University, Karaikudi.*

*(Corresponding author: C. Gnana Chithra)*

*(Received 28 September, 2016 Accepted 29 October, 2016)*
*(Published by Research Trend, Website: www.researchtrend.net)*

**ABSTRACT: This paper explores the new methodology of detecting boundaries of the sentence by heuristic method and also classifies it. Automatic true detection of the sentence aids in semantically annotating the web. Sentences formed with URL, ellipsis and abbreviations are focus of the study. High performance features are selected for Classification using C4.5 decision trees and K-Means for clustering with the help of datasets. Sentences Classified by human annotators, Manning's Heuristic algorithm, the proposed Modified Manning's algorithm, and machine learning supervised and unsupervised algorithms are evaluated. Heuristic learning adapted by this system produces an average F1 score of 96.58% for SBD.**

**Keywords**: SBD, Sentence Boundary Detection, C4.5, K-Means, Heuristic SBD.

## I. INTRODUCTION

Oxford dictionary [1] defines Sentence as "A set of words that is complete in itself, typically containing a subject and predicate, conveying a statement, question, exclamation, or command, and consisting of a main clause and sometimes one or more subordinate clause." Sentence Boundary detection is the most preliminary pre-processing step in Natural Language processing such as Named Entity Recognition, POS tagging, Information Extraction, Information Retrieval, Automatic summarization, Discourse analysis, Machine Translation, Morphological Segmentation and Automatic Speech Recognition systems. This area of research has grabbed recent attention to achieve the goal of Semantic web by enhancing the speech recognition outputs in the ASR's and in language processing modules.

A methodology, which identifies the clear disambiguous Sentence Boundary detection, could guide in creation of best ontology and well-defined corpus. Since all the current algorithms cannot handle many special cases such as abbreviations, URL, colon, imbalanced parenthesis and brackets, the proposed algorithm handles the same.

Section 2 presents reviews of the related work.Rule based learning and machine based learning is described in section 3. Features for Sentence boundary detection is discussed in Section 4. Section 5 deals with datasets. Section 6 is dedicated to the performance evaluation of the results and Section 7 presents the conclusion.

## II. RELATED STUDIES

Riley [2] initially proposed the machine learning strategy-using feature set for testing the incidence of periods in the sentence boundary detection. His method yielded result of 99.8% accuracy.

Reynar and Rathnaparki [3] applied Maximum entropy to classify the sentence boundary detection. Their algorithm took into consideration only the prefix and suffix of the End of sentence marker. They also proved that on addition of words to the context did not help further. This Maximum entropy model yielded accuracy of 98.8% for domain-dependent model and 98% for domain-independent model.

Kiss and Strunk [4] proposed the Punkt Sentence Tokenizer, which identifies the abbreviation, based on the unsupervised approach. His accuracy factor on WSJ data was 98.98%

Sentence boundary detection was not only limited to English language. Yuya Akita.et.al [5] has applied SVM (Support Vector Machine) and SLM (Statistical language model) to find sentence boundaries in Japanese.

Grefenstette and Tapanainen [6] experimented rule based classification with regular expression to differentiate the occurrence of period in email, web addresses and numbers. It reported accuracy of 93.78%.

Gillick [7] proposed a statistical system using Support vector machine for learning and full stops as sentence boundaries with a success rate of 99.75% in WSJ corpus.

## III. METHODOLOGY

### *A. Definition of Sentence Boundaries*

Sentence Boundary can be simply stated as an end of the first sentence and the beginning of the second sentence. A clause may be considered as a cluster of words with subject and predicate. Independent clause can be a complete sentence due to its fullness whereas Dependent clause cannot survive by itself but depends on the Independent clause. Sentences can be classified into four types based on their nature. The four types of boundaries viz. Supreme boundary, Robust boundary, fragile boundary and frail boundary very well depend on the wholeness of the syntax clause as well as their semantic means.

Supreme boundary can be defined as a complete sentence. Robust boundary are assumed to be Strong boundaries due to their Independent clauses. The fragile and frail boundaries are determined on the degree of the conditional clauses used. It is difficult for the AI system to determine the fragile and frail boundaries and the help of human expert annotators provides way for the solution.

### *Manual Classification of Sentence Boundary detection*

To obtain the Gold data, six English Language experts in the Master's level education were assigned the task of punctuating the sentence and marking the sentence boundary in the document where punctuation's were stripped off. Each was given 10 documents from different domain containing 3500 sentences and approximately 32000 words. Experts normally do not agree among themselves in punctuating sentence. The linguists were all given the same text, but the results proved to be different. There were discrepancies in identification of sentence boundaries. The results of the classification are shown in Table 1.

**Table 1: Evaluation of Manual Sentence Boundary detection.**

| Experts | Precision | Recall | F-Measure |
|---|---|---|---|
| Expert 1 | 94.29% | 96.35% | 95.30% |
| Expert 2 | 98.47% | 97.92% | 98.19% |
| Expert 3 | 99.32% | 98.00% | 98.65% |
| Expert 4 | 97.64% | 98.23% | 97.93% |
| Expert 5 | 99.70% | 96.65% | 98.15% |
| Expert 6 | 94.40% | 97.37% | 95.85% |

Precision and Recall is one of the classical evaluation metric in the field of IR.It is the calculation of True positives, false positives and False Negatives. In our case Precision is the correct percentage of EOS (Period) marker proposed by the experts while recall is the percentage of EOS that happens to be in the corpus. Investigations conducted by Beeferman et.al. [8] explains the fact that humans never agree unanimously on the acceptability of inserting commas and their agreement on Sentence boundaries.

### *Sentence Segmentation*

After the document is extracted from the web, the sentence should be segmented with the boundaries to obtain the tokens. There should not be any ambiguity in boundary detection. A sentence usually ends with the delimiter full stop (.) OR Question mark [?] or Exclamatory mark (!). Paragraphs are made of sentences. Sometimes it is very challenging to extract the sentences due to the ambiguity in the punctuation markers. When the Sentences are within quotes it cannot be identified easily. Rule based learning uses regular expression for sentence segmentation. Sentence markers must obey some sort of regularity in the sentence. Otherwise it is still a challenge to break a sentence from the paragraph. The Linguistic analysis made in this work is unique to the English character set.

### *Sentence Segmentation Heuristics*

Heuristic 1(Heu1): The period character '.' in the name of the initials of a person should not be split into a separate sentence. For Example A.R.Rahman is a music Composer and song Writer. The sentence should be not split after A or R because those are initials.

Heuristic 2 (Heu2): The period character in the name of educational Degrees should not be spilt into sentences. For e.g. Post graduation Degree on Physical education M.P.Ed gives rich knowledge on theory as well practical skills.

Heuristic 3 (Heu3): The URL should not be split as it contains periods.For E.g.https://www.crisilresearch.com contains Economic and Financial research reports of the Sectors as well as individual companies.

Heuristic 4 (Heu4): When there are names after abbreviation it should not be split.

Mt. Everest is the highest Peak in the world. Mt stands for mountain and Everest is a Proper noun.

Heuristic 5 (Heu5): Sentence should not be split after Ellipses in English. For E.g. On hearing about the Earthquake, Mr. Obama said, "Oh God! I do not know what to do …" but Vice President said we can manage the situation with the existing resource.

Heuristic 6 (Heu6): When there is an imbalance in the parenthesis or bracket of sentence do not split the sentence. "The population in Africa is 1.111 billion (2013}. When economic development is concerned it is far behind other continents".

*Proposed system*

A new heuristic algorithm was developed based on the Sentence boundary detection algorithm by Manning et.al. [9]. This rule based heuristic system automatically identifies the boundaries of the sentences with End-of-Sentence marker. Abbreviation poses a major problem in Segmentation. To solve this ontologies are used.

Though ontology does not provide a pure virgin concept model independent of domain, but provides domain dependent rich set of concepts, which improves the optimality of the solution. Word net ontology, Geographical Gazetteer and University Degree and Diploma ontology are integrated to find out the abbreviations used in the sentence.

*Manning's Algorithm for Sentence Boundary Detection*

Manning's algorithm places the EOS marker after all the occurrences of. ?! ; : and -. If any quotation marks are encountered the boundary is moved. Since the period (.) is assumed as boundary sometimes it is disqualified due to situations such as a title given to the name, which may be family titles, professional title, or honorary titles when preceding the name may interlude. The boundary is discarded after an? Or ! followed by a lower case letter. n general about 90% of periods are sentence boundary indicators. (Riley 1989). Modified Manning's Heuristic Sentence Boundary Detection algorithm is based on the Heuristic Sentence Division algorithm by Manning.et.al.

**Manning's Heuristic Sentence Division algorithm**

- ➢ Place putative sentence boundaries after all occurrences of . ? ! (and maybe ; : -_)
- ➢ Move the boundary after following quotation marks, if any.
- ➢ Disqualify a period boundary in the following circumstances:
    - If it is preceded by a known abbreviation of a sort that does not normally occur word finally, but is commonly followed by a capitalized proper name, such as Prof. or vs.
    - If it is preceded by a known abbreviation and not followed by an uppercase word. This will deal correctly with most usages of abbreviations like *etc.* or Jr. which can occur sentence medially or finally.
- ➢ Disqualify a boundary with a? or ! if:
- ➢ It is followed by a lowercase letter (or a known name).
- ➢ Regard other putative sentence boundaries as sentence boundaries.

**Fig. 1.** Heuristic Sentence Division Algorithm by Manning et.al.

**Reprinted from "Foundations of statistical natural language processing "by Manning, C.D. and. Schütze., H. 2002. The MIT Press, London.**

*Modified Manning algorithm for SBD*

**MODIFIED MANNING'S HEURSITIC ALGORITHM**

- ➢ Place putative sentence boundaries after all occurrences of. ? ! (and maybe ; : -_)
- ➢ Move the boundary after following quotation marks, if any.
- ➢ Disqualify a period boundary in the following circumstances:
    - If it is preceded by a known abbreviation of a sort that does not normally occur word finally, but is commonly followed by a capitalized proper name, such as Prof. or vs.
    - The period character '.' in the name of the initials of a person should not be split into a separate sentence.
    - The period character in the name of educational Degrees should not be spilt into sentences.
    - Lookup the ontology for recognizing the educational qualification.
    - If Abbreviation contains numbers check it against the ontology.
    - Abbreviations other than educational degrees and geographical data are referred with Wordnet ontology and ontology containing honorary titles, family titles and professional titles.
    - The URL should not be split as it contains periods.
    - Sentence should not be split after Ellipses in English.
- ➢ Disqualify a boundary with a ? or ! if:
- ➢ It is followed by a lowercase letter (or a known name).
- ➢ When there is an imbalance in the parenthesis or bracket of sentence, do not split the sentence. Balance the parenthesis or bracket by inserting or replacing the mark.
- ➢ Regard other putative sentence boundaries as sentence boundaries.

**Fig. 2.** New proposed algorithm "Modified Manning's Heuristic Algorithm".

**Partially Adapted from"Foundations of statistical natural language processing " by Manning, C.D. and. Schütze., H. 2002. The MIT Press, London.**

Manning's algorithm has been modified for the purpose of efficient Sentence boundary detection. The algorithm can be extended such that is takes of abbreviations and does not split the sentence after abbreviation. Latest revolution technology such as Ontology engineering can be used for SBD.The titles such as academic title, Fellow of an artistic or Professional body, Military title of honor needs to be identified and the sentence should not be split after the titles.

For e.g. “ Indian NSG officer Lt.Col. Niranjan Kumar, martyred in Pathankot terrorist attacks was appreciated by the entire Indian Nation for his bravery”. Here Lt.Col. is a Military honor and it is different for different countries. Occasionally period falls in-between the honor and the name of the person.For e.g. “Bharat Ratna Dr. M.G.Ramachandran was the former Chief minister of Tamilnadu”. Here Bharat Ratna is a honorary title followed by Dr. In this case EOS marker should be placed only after Tamilnadu.

Geographical abbreviations need to be clearly marked to avoid the wrong EOS.

Mt. Kanchenjunga, is the third largest mountain and it lies in Nepal and Sikkim.

River Ganges can also be written as R. Ganges. Gazetteers help in clear identification of Geographic shorthand notations. Abbreviations may sometimes contain numbers. For e.g. G-20 was held in Turkey during 2015. The hyphen in G-20 cannot be mistaken for sentence boundary because G-20 is the name of the summit in total. So ontology lookup is essential for abbreviations with numbers.

Using hand crafted rules the URL’s are identified and their boundaries can be detected without any ambiguity and ellipsis should be classified not as sentence boundary.

*Machine Learning algorithms for SBD*

In the field of Data mining, Machine Learning algorithms are used for learning the accurate predictions with the help of past data and classifying the data into categories. The dataset is divided into training data and test data. Training data contains the classified examples where as test data is the one that needs to be classified based on training data. Machine learning algorithms are cheap when compared to human classification; as well it can handle large scale of data. Accuracy, Speed, Robustness and Scalability are the salient features of machine learning algorithms.

**Sentence Boundary Detection Using Machine learning algorithms**

1. **K-means clustering**
2. **C4.5classifier**

**1) K means Clustering**

K-means[10] is an unsupervised machine-learning algorithm, which solves the problem of clusters. Clustering is an mechanism, which splits a collection of data points into a certain small number of clusters. K – centroids should be defined for all the clusters. The centroid should be placed far away from each other as it produces different results when placed in various other areas.

Each and every data point in the dataset has to be linked to the nearest centroid until all the data points are associated. In this stage first step of grouping is done. Again we calculate K-new centroids and link the data points to the newly found centroids. A loop begins to form and the K-centroids keep on changing their position until there is no shift of movement of the centroid. Finally the K-means algorithm converges to a point, which may not be, squared error function. The clustering stops when the binding of data points to the centroids does change in successive iterations. The goal of the K-means algorithm is the minimize the squared error function.

**Advantages and disadvantages**

1. K-Means is a very simple algorithm.
2. When the clusters are kept small, although the feature vector is large, its computation performance is faster than other clustering methods.
3. Algorithm can produce stronger clusters when compared to Hierarchical clustering.

**Disadvantages**

1. The model of cluster is the major limitation. Only spherical clusters are taken into account for easy partitioning.
2. It poses difficulty to predict the k-centroid.
3. The performance of the algorithm varies strongly with the clusters of varied size.
4. The final cluster output is dependent on the initial selection of the partition.
5. It works well for some dataset and poor visibility of clusters on others.

**2) C4.5 Classifier**

C4.5 [11] is a classification algorithm, which outputs classification rules in the form of decision tree, and it is a solid base for many algorithms. This algorithm developed by Ross Quinlan in 1993 was an extension of ID3 algorithm. C4.5 statistical classifier uses the Information gain for splitting samples. Information gain is helpful for measuring the gain ratio. The best feature of this algorithm is that it can handle continuous, discrete and missing values. Threshold level is defined for continuous values and the attributes are split into two, one above the threshold level and the other which is less than or equal to the threshold level.

Let us assume that TI is the set of training instances. Choose a attribute A1 from TI. Select the initial subset of training data S. The decision tree is built on the attribute A1 and S. Setting aside the subset S consider the rest of data for testing the integrity and accuracy of the decision tree. Find out if all the data is correctly classified or not. When correctly classified and when the data is pure and the stopping criterion is met then stop the further classification. Otherwise add all the wrong instances to the initial subset S and build a new decision tree. Iteration is carried out until a decision tree is built on all the data that has been classified correctly.

During the construction of Decision tree, at times missing values or unknown data is encountered. The missing data for that attribute is calculated using the gain ratio.

Sometimes over fitting the data occurs and proves to be an important issue in decision trees. The prediction rate also decreases. Pruning is a machine learning technique, which can reduce the prediction error rate. Pruning cuts the size of Decision tree by trimming the branches of the tree, which are less important in the classification process.

**Advantages**

1. Computational implementation of this machine-learning algorithm is easy.
2. Continuous and discrete or categorical values can be handled by C4.5, by defining the threshold value and splitting the dataset into two.
3. The prototypes created by this algorithm can be understood very easily.
4. Missing data or noisy data are dealt by this algorithm.
5. Can build smaller or larger accurate Decision Trees.
6. Best methods for pruning of trees.

**Disadvantages**

1. The major limitation is, it performs well only with larger training sets. On the contrary it performs worst with small training sets. It is very sensitive to values.
2. Creation of empty branches by C 4.5 poses a great problem.

*Features for SBD*

SBD problem can be regarded as a classification task, which means that all the word in the text is analyzed for classification. It is a clustering task as well. If the word is a sentence boundary then it is categorized into SENTENCE_BOUNDARY CLASS or otherwise NO_SENTENCE_BOUNDARY CLASS.We can call the Present word as Current, the word following the current word is the Next, and word preceding the current word is Previous.

In this research work, added to fourteen features that are adapted from Neha Aggarwal et.al. [12], we introduce eight more features to detect the sentence boundary. In total the classifier uses these feature set to classify a word as a sentence boundary or not.

The feature set focus more on the use of ellipsis, triple exclamatory, tri question mark, abbreviations and URL handling.

DATASET

Four Datasets was prepared with great attention for its coverage on large set of samples for the purpose of training and testing the Sentence Boundary Detection. Different domains are considered for this study. News, Tourism, Geographic, Food, Financial, Blogs, and Military are some examples of domains that was investigated for this study. Some data was also extracted from Wall Street Journal dataset, Brown corpus and Reuters. The size of the dataset is not all similar. Dataset (DS-1) was extracted from the Reuters corpus, Wall Street Journal and News domain. Dataset (DS-2) contains the partial data to fit the memory requirement of the testing system from the food and novel domain. Dataset (DS-3) has few data from Brown corpus; financial pages domain and military domain and Dataset (DS-4) from the tourism and geography domain. Different cases for abbreviations, urls and ellipsis are placed in the dataset. DS-1, DS-2, DS-3, DS-4 contains 5000, 4000, 2000, 1000 sentences respectively.

**Table 2. List of features for SBD.**

| |
|---|
| Previous/Next is Uppercase |
| Previous in all Uppercase |
| Previous/Next length |
| *Current* is “:”,”--”,”…” |
| *Next* is “$” |
| *Next* is all digits |
| *Previous* is an abbreviation |
| *Current* is “.”,”?”, or “!” and *Next* is “--” Or double left quote (“) |
| *Current* is “.”,”?”, or “!” and *Next* is Not double left quote |
| *Previous* is “.”,”?”, or “!”, *Current* is single or double right quote (_ or ”), and *Next* is double left quote (“) or is uppercase |
| *Current* is uppercase |
| *Previous* is “.”, “?” or “!” |
| *Current* is “.”, “?”, “!”, “!!”, “??”, “-RRB-“, single quote, double quote, or double right quote<br>*Next* is “-LRB-”, single quote, double quote, or double right quote |
| Current is ellipsis(…) |
| Current is triple exclamatory(!!!) |
| Current is di exclamatory(!!) |
| Current is tri question mark(???) |
| Previous is word, current is period(.), Next is Proper noun |
| Previous is Abbreviation, current is period (.), Next is Proper noun |
| Abbreviation containing hyphen (-) |
| Current is http or https or www and Next is “/” |

Note: Agarwal N., Ford K., and Shneider M., “Sentence Boundary Detection Using a MaxEnt Classifier,”*in Proceedings of MISC*, CA, pp. 1-6, 2005.

EVALUATION

In order to obtain Gold Standard dataset, six experts in English from expertise in different domains were asked to punctuate the sentence and find the End-of-Sentence (EOS) markers. There was a slight variation in the Inter Expert Agreement (IEA), but within the tolerable limits. Results on evaluation of IEA on different datasets are given in Table 3. Apart from the 4 datasets they were provided with 3500 sentences for multiple domains to obtain training data.

Though period (.) is the common end of sentence marker, other EOS markers were also classified. While human evaluation, built-in knowledge on Parts of Speech helped to punctuate the datasets more clearly.

**Table 3. Evaluation of IEA on datasets.**

| Dataset | Inter Expert Agreement |
|---|---|
| DS-1 | 0.98 |
| DS-2 | 0.97 |
| DS-3 | 0.98 |
| DS-4 | 0.99 |

Detailed analysis was also performed on the End of sentence markers in a document. Results of the analysis are recorded in Table 4. The major slice of percentage of EOS markers is obtained by the period (.). Though 82.10% is the EOS marker percentage, in Total we get 91.23%, which acts as a full stop. Though Tri-Exclamatory, Tri-Question mark, Ellipsis in total makes a small slice as EOS tag, it is very important to achieve high results and fulfill the dream of semantic web.

**Table 4: Investigation on End of Sentence markers.**

| EOS markers | Percentage of EOS markers detected as sentence Boundaries |
|---|---|
| Period (.) | 82.10% |
| Question Mark (?) | 6.40% |
| Exclamatory Mark (!) | 3.07% |
| Colon (:) | 5.32% |
| Tri Exclamatory (!!!) | 0.40% |
| Tri Question mark (???) | 0.15% |
| Ellipsis (…) | 0.36% |
| Miscellaneous | 2.20% |

Precision also called as Positive Predictive Value (PPV) measures how much of the retrieved documents during the search are relevant. It is the basic means of evaluating a search algorithm. Precision can be defined as the ratio of the retrieval of relevant records to that of total number of relevant and irrelevant recordsduring its retrieval. The Precision values on the four datasets by four different algorithms are given in Table 5. In the Dataset DS-1 K-means algorithm has the highest precision of 96.52% and Modified algorithm has 95.23%. Due to the noisy data algorithms except C4.5 could not perform better. In all the four datasets K-means performs best on the precision and modified algorithm is better than its predecessor. The graphical representation of Precision, Recall and F-measure evaluation is given inFigure 3, Figure 4, Figure 5 and Figure 6.

The Mean Average Precision (MAP) can also be measured against the recall to check the efficiency of the algorithms.

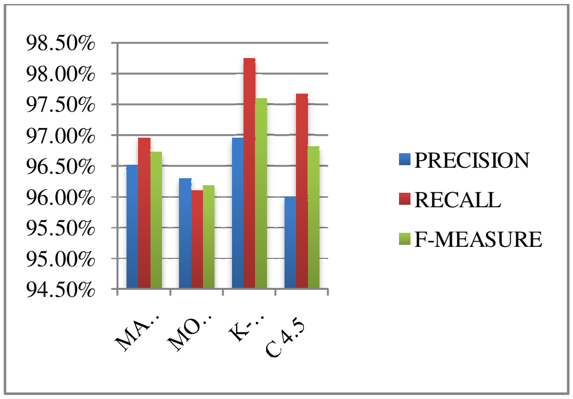

**Fig. 3.** Cluster chart for Precision, Recall, F-measure on DS-1.

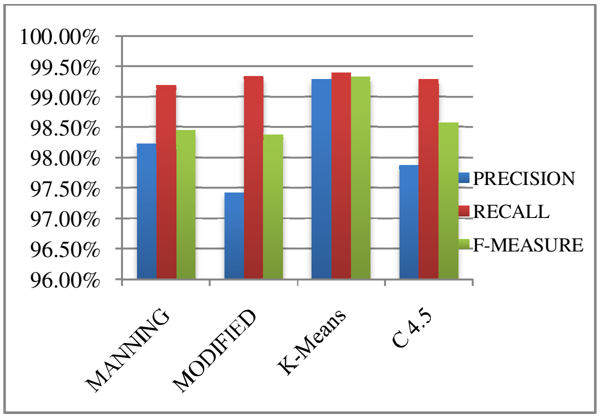

**Fig. 4.** Cluster chart for Precision, Recall, F-measure on DS-2.

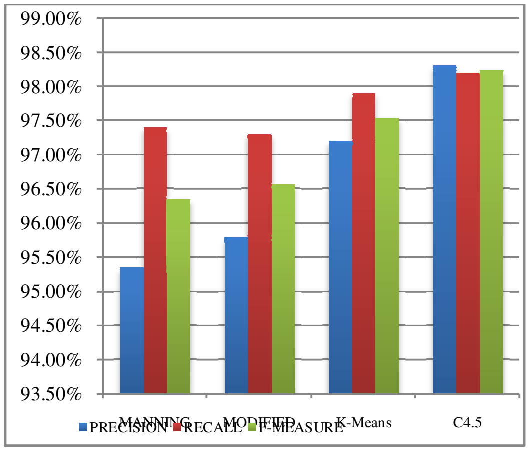

**Fig. 5.** Cluster chart for Precision, Recall, F-measure on DS-3.

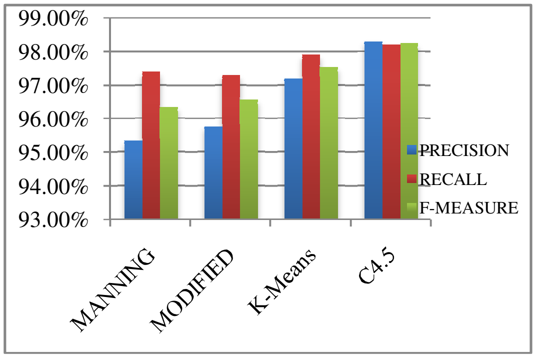

**Fig. 6.** Cluster chart for Precision, Recall, F-measure on DS-4.

**Table 5. Comparative PRECISION values by algorithms on Datasets.**

| Dataset | Precision | | | |
|---|---|---|---|---|
| | Mannings Algorithm | Modified Algorithm | K-Means | C 4.5 |
| DS-1 | 94.31% | 95.23% | 96.52% | 95.70% |
| DS-2 | 96.52% | 96.30% | 96.96% | 96.00% |
| DS-3 | 98.23% | 97.43% | 99.30% | 97.89% |
| DS-4 | 95.34% | 95.78% | 97.20% | 98.3% |

Recall also called as Sensitivity or as a measure of retention. It can be defined as the ratio of the number of relevant record found to the total number of relevant records during the search. In Table 6. DS-3 has recorded a high recall rate above 99%.

Apart from the machine learning algorithms, rule based algorithms has proved to possess a good recall rate with an average of97% for modified algorithm.

**Table 6: Comparative RECALL values by algorithms on Datasets.**

| Dataset | Recall | | | |
|---|---|---|---|---|
| | Manning's algorithm | Modified algorithm | K-means | C 4.5 |
| DS-1 | 95.26% | 95.15% | 97.13% | 98.2% |
| DS-2 | 96.95% | 96.10% | 98.24% | 97.67% |
| DS-3 | 99.20% | 99.35% | 99.40% | 99.30% |
| DS-4 | 97.40% | 97.30% | 97.90% | 98.20% |

F-measure tells how precise the classifier is and its robustness. Modified algorithm has 95.18% on DS-1, 96.19% on DS-2, 98.38% on DS-3 and 96.57% on DS-4. F-measure on DS-3 is the overall high performer. In the K-means algorithm F-measure is 99.34% on DS-3 and in C4.5 decision tree it is 96.93 in DS-1, 96.82% in DS-2, 98.58% in DS-3 and 98.24% in DS-4.

Overall the modified algorithm can be classified as best in rule-based mining due to the high Precision, Recall and F-measure in that segment.

**Table 7. Comparative F-MEASURE values by algorithms on Datasets.**

| Dataset | F-measure | | | |
|---|---|---|---|---|
| | Mannings algorithm | Modified algorithm | K-means | C 4.5 |
| DS-1 | 94.78% | 95.18% | 96.82% | 96.93% |
| DS-2 | 96.73% | 96.19% | 97.59% | 96.82% |
| DS-3 | 98.46% | 98.38% | 99.34% | 98.58% |
| DS-4 | 96.35% | 96.57% | 97.54% | 98.24% |

The graphical interpretation for the Precision, Recall and F-measure by four different algorithms on four independent datasets is represented in Figure 3, Figure 4, Figure 5, and Figure 6.

## IV. CONCLUSION

This paper deals with different approaches for Sentence Boundary Detection problem. They were Rule based and Machine learning based. C4.5 Decision tree Supervised algorithm and K-Means Unsupervised algorithm were used. First, Manning's rule based Sentence division algorithm was altered with new criterion as Modified Sentence Boundary detection algorithm, which achieved a remarkable F measure of 96.58%

Since ontologies and Gazetteers were used for look up, rule based mining came up with excellent results. Detecting Abbreviations and Geographical entities is the highlight of this algorithm. C4.5 classifier used the best feature set to produce the outstanding results. The predictive accuracy, speed, interpretability, robustness and scalability of the classifier were good. K-means clustering algorithm used small clusters to segregate the data into Sentence Class and No-Sentence Class. Though K means algorithm was classified as Unsupervised it used some training data for clustering. Hence it can be also termed as semi supervised algorithm.

This Work can be extended to the sentences in Social networking sites such as Facebook and Twitter. More features for automatic sentence boundary detection can be learnt from Social media. It can also be extended to Automatic Speech Recognition systems.

## REFERENCE

[1]. https://en.oxforddictionaries.com/definition/sentence

[2]. Michael D. Riley. 1989. Some applications of tree based modelling to speech and language. In Proceedings of the Workshop on Speech and Natural Language, HLT '89, pages 339–352, Stroudsburg, PA, USA. Association for Computational Linguistics.

[3] Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In Proceedings of the 5th Conference onApplied Natural Language Processing, pages 803–806, Washington, DC, April. ACL.

[4] T. Kiss and Strunk, J. 2006. Unsupervised multilingual sentence boundary detection. Computational Linguistics, 32(4):485–525.

[5] Akita, Y. 2006. Sentence Boundary Detection of Spontaneous Japanese Using Statistical Language Model and Support Vector Machines. In: Proceedings of. Interspeech-ICSLP, Pittsburgh, PA.

[6] Gregory Grefenstette and Pasi Tapanainen. 1994. What is a word, what is a sentence? problems of tokenization. In Proceedings of the 3rd Conferenceon Computational Lexicography and Text Research, Budapest, Hungary, July.

[7] Dan Gillick. 2009. Sentence boundary detection and the problem with the u.s. In Proceedings of Human Language Technologies: The 2009 Annual Conferenceof the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, NAACL-Short'09, pages 241–244, Stroudsburg, PA, USA. Association for Computational Linguistics.

[8] Beeferman, A. Berger, and J. Lafferty. 1998. CYBERPUNC: A lightweight punctuation annotation system for speech. In *Proceedings of the IEEEInternational Conference on Acoustics, Speechand Signal Processing,* pages 689-692, Seattle,WA.

[9] Manning, C.D. and. Schütze., H. 2002. Foundations of statistical natural language processing. The MIT Press, London.

[10] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

[11] J.R. Quinlan, C4.5 programs for machine learningMorgan Kaufmann Publishers, livres Google.

[12] Agarwal Neha, Kelley Herndon Ford, Max Shneider, "Sentence Boundary Detection Using a MaxEnt Classifier, "*in Proceedings of MISC*, CA, pp. 1-6, 2005.Biblio for reuters dataset, brown dataset, wsj