THE MACHINE LEARNING METHOD REGARDING EFFICIENT SOFT COMPUTING AND ICT USING SVM

1TARUN DHAR DIWAN, 2UPASANA SINHA
ASSISTANT PROFESSOR
DEPT. OF ENGINEERING

1Dr.C.V.RAMAN UNIVERSITY, BILASPUR (INDIA)
2J K INSTITUTE OF ENGINEERING, BILASPUR (INDIA)
1taruncsit@gmail.com, 2upasana.sihna@gmail.com

ABSTRACT

Soft Computing and Information Communication Technology (ICT) both are main part in information technology and computer science because today all documentations are computerized to store necessary information and database into English language, so natural language processing concept is come from artificial intelligence (A.I) branch of computer science and information technology as part of soft computing and ICT (information communication technology). Here English language may be used as a major part of machine learning method, where sentence boundary detection is a great challenge in current time of NLP(natural language processing). So SVM(support vector machine) can be used to solve the problem of sentence boundary detection which indicates the machine learning method regarding efficient soft computing and ICT. Hence Computer system may be learned to avoid ambiguity of dot (.) periods in detecting sentence boundary.

I. INTRODUCTION

Basically sentence boundary detection is essential feature of recognizing the correct sentence meaning in text area and natural language processing, because without valid meaning based documentation has no any information extraction features and communication is generated in efficient successful trend. SVM (support vector machine) may be used to extract the feature of sentence boundaries where Support Vector Machine (SVM) is a machine learning model and which solved the problem of sentence boundary detection by calculation of precision, recall, f feature value [1]. Feature extraction f value is evaluated on
the basis of precision, recall value calculation. Sentence boundary detection (SBD) means detecting end of sentence in given or used text. Various type of punctuation marks contain multiple feature at one time which gives sentence boundary ambiguity in text. This type of error of ambiguity can be find by calculating precision, recall values and feature values of support vector machine and in such a way sentence boundary is detected, where ambiguity is come due to multiple behavior of one punctuation mark or symbols in text at one time, then precision and recall values are calculated to provide better information on what kinds of errors were made by ambiguous punctuation mark. Basically SVM (support vector machine) firstly calculate ambiguity of each segment in input English text by calculating precision value, here segment is one part of input text but not lead to exact sentence boundary [2]. Now we calculate recall value to find exact sentence boundary and $f$ is calculated at last to extract features of each segment of input text.

II. RELATED WORK

Basically here sentence boundary detection is done on the basis of dependency analysis of each segment of text. Segment of text is not necessary to occur as a boundary of segment so there dependency analysis of previous and next segment of text is done in three type of sentence which is given bellow:

1. Open dependency based sentence.
2. Closed dependency based sentence.
3. Without dependency based sentence i.e. Independent sentence.

In above three type of sentence based ambiguous text has a lot of need of detecting sentence boundary detection and after analysis of this three type of sentence in this paper by using suitable example, it is search that more feature of text segment gives more ambiguity with minimum SBD (Sentence Boundary Detection) [1,3].

III. CHALLENGES/ PROBLEMS OF SENTENCE BOUNDARY DETECTION

In sentence boundary detection following problems are come as challenges to detect end of sentence in creating recognize correct meaning of any text of English language during soft computing or information communication technology (ICT) [4].

1. Recognize tokens by removing white space and special characters from English text.
2. Resolving ambiguous separators in numbers to text.
3. Resolving ambiguous abbreviation in text.
4. Then resolving end of sentences (EOS) (?).
5. Morphology analysis on a corpus database in a form of statistical analysis and word/letter analysis which are part of NLP.
6. Correct syntactic and semantic meaning of each word by using grammar i.e. morphology and lexical analysis.
7. Avoiding ambiguity in punctuation marks.
IV. OBJECTIVE/GOAL OF PAPER

Finding the solution of sentence boundary detection challenge as avoiding ambiguity of dot(.) punctuation mark. It is done by calculating the error rate where error rate is related to each segment of text which is evaluated to check dependency analysis on the basis of dot(.) Punctuation as either it is a abbreviation, termination or name into English language based text.

V. SIGNIFICANT OF PAPER

In this paper ambiguity is avoided of dot(.) periods into each segment of English language based text as to detect sentence boundary where f feature value is detected from SVM as when any word is presence it is counted as 1 otherwise as 0 value and this 0 or 1 value is assign finally assign into vector of SVMs.

VI. PROPOSED METHODOLOGY

SVM methodology is support the creation of document vector from the features extraction value of each segment or line of text. Each new word seen by the SVM module is internally assigned to a different coordinate in the vector. The value of each coordinate is zero when the word is absent from a document otherwise one, basically it is a binary arrangement of text segment vector in memory[5][6].So maximum feature f value is created with maximum presence of one word at one time into one segment of text, which represent high maximum ambiguity i.e. maximum precision value, and maximum ambiguity of any word or string into one segment or line of text ,decreases the sentence break or boundary detection behavior as a less(minimum)recall value. This is clear by using bellow example of input, output text along precision, recall, f feature extraction value. To apply methodology, calculation is done to calculate true positives, false positives, true negatives, and false negatives values whose description is bellow given. These values are used in recall, precision, f feature extraction value calculation formula[7][8].

1. True Positives: Those who test positive for a condition and are positive (i.e., have the condition).
2. False Positives: Those who test positive, but are negative (i.e., do not have the condition).
3. True Negatives: Those who test negative and are negative,
4. False Negatives: Those who test negative, but are positive.

Now Precision is the ratio between the number of candidate tokens that have been correctly assigned to a class and the number of all candidates that have been assigned to this class [9][10].
Precision value=true positives / (true positives + false positives) ------- (1)

Recall is defined as the proportion of all candidates truly belonging to a certain class that have also been assigned to that class by the evaluated system.
Recall value=true positives / (true positives + false negatives) ------- (2)
Finally, the so-called F measure is the harmonic mean of precision and recall (van Rijsbergen 1979).

\[
\text{f measure value} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad ---------(3)
\]

VII. ANALYSIS

Here if any sentence is closed with proper syntactic and semantic analysis where start and end both are related to only one sentence then this type of sentence is come into without dependency or independent type of sentence, but if end is correct as syntactic and semantic way but starting is depend on previous segment of text then this type of sentence is known as closed dependency based sentence and if both starting and ending of any segment of text are incomplete the this type of segment is known as open dependency based sentence.

Therefore sentence boundary is introduced as three types on the basis of above description of analysis i.e.
1. Strong boundary(independent sentence i.e. without dependency based sentence)
2. Weak boundary(open dependency based sentence)
3. Absolute boundary(closed dependency based sentence)

Finding Dependency Structure of Each Sentence to Used Input Text

Example of Input English Text
I am doing Ph.D. Ph.D. is easy. Mr. and Mrs. Sharma are well. They are Mr. and Mrs. Arun is singing a song. Our college name is S.E.C. Mr. V. Dixit is a best guide.

Expected Output after Used Methodology to Input English Text
I am doing Ph.D.
Ph.D. is easy.
Mr. and Mrs. Sharma are well.
They are Mr. and Mrs.
Arun is singing a song.
Our college name is S.E.C.
Mr. V. Dixit is a best guide.

Finding Four Type of Sentence Dependency Structure
1) I am doing Ph.D. Ph.D. is easy. Mr. and Mrs. 
   (Open dependency based sentence)
2) Sharma are well. They are Mr. and Mrs. Arun is
   (Open dependency based sentence)
3) playing a game. Our college name is S.E.C.
   (Closed dependency based sentence)
4) Mr. V. Dixit is a best guide.
   (Without dependency based sentence i.e. independent sentence)
Here Precision to each statement.
10/10+7=1.4=10/17=0.588
10/10+3=3.3=10/13=0.769
8/8+6=1.3=8/14=4/7=0.571
7/7+3=2.3=7/10=0.7

Here Recall to each statement.
10/10+2=5=10/12=5/6=0.8333
10/10+2=5=10/12=5/6=0.8333
8/8+1=8=8/9=0.888
7/7+1=7/8=0.875

&

So f measure values-

\[
\frac{2(0.588 \times 0.833)}{(0.588 + 0.833)} = 0.688 \\
\frac{2(0.769 \times 0.833)}{(0.769 + 0.833)} = 1.602 \\
\frac{2(0.571 \times 0.888)}{(0.571 + 0.888)} = 0.694 \\
\frac{2(0.875 \times 0.7)}{(0.875 + 0.7)} = 0.777
\]

Here input English text has a evaluation of four type of sentences i.e. first two type are open dependency based sentence, third closed dependency based sentence and finally last fourth type of sentence is without dependency based sentence and above calculation proof that max precision gives max ambiguity regarding to abbreviation, punctuation, separator so sentence boundary has a need of evaluating ambiguity and then eliminating it to gain exact sentence boundary to any input text.

Sentence boundary maximally detected to open dependency structure based text because features are maximum with various individual sentence boundaries.

VIII. RESULT

Here f is feature measure i.e. lexical analysis, parsing or morphological analysis.
Recall is exact sentence boundary detection value calculation.
Precision is ambiguity calculation of each segment of used text.
Hence feature extraction value f is max 1.6% in open dependency based sentence to maximum ambiguity of 0.8% precision value and corresponding exact sentence boundary recall value is minimum i.e. 0.83%.
<table>
<thead>
<tr>
<th>Sno</th>
<th>Type of Sentence</th>
<th>Recall Value</th>
<th>Precision Value</th>
<th>F measure Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>With dependency information (open)</td>
<td>10/10=1</td>
<td>10/10=1</td>
<td>2* .588** .633** (.588** .633** )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10/12=8/12</td>
<td>10/12=8/12</td>
<td>= .979/1.421</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5/6=0.833</td>
<td>5/6=0.833</td>
<td>= 0.688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=0.83%</td>
<td>=0.83%</td>
<td>=0.7%</td>
</tr>
<tr>
<td>2</td>
<td>With dependency information (open)</td>
<td>10/10=1</td>
<td>10/10=1</td>
<td>2* .769** .833** (.769** .833** )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10/10=1</td>
<td>10/10=1</td>
<td>=1.602</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=0.8%</td>
<td>=0.8%</td>
<td>=1.6%</td>
</tr>
<tr>
<td>3</td>
<td>With dependency Information (closed)</td>
<td>8/8=1</td>
<td>8/8=1</td>
<td>2* (.571** .888** )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8/9=0.888</td>
<td>8/9=0.888</td>
<td>=1.014/1.459</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=0.89%</td>
<td>=0.89%</td>
<td>=0.694</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=0.6%</td>
<td>=0.6%</td>
<td>=0.7%</td>
</tr>
<tr>
<td>4</td>
<td>Without dependency Information</td>
<td>7/7=1</td>
<td>7/7=1</td>
<td>2* (.875** .750** )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7/8=0.875</td>
<td>7/8=0.875</td>
<td>=1.225/1.575</td>
</tr>
<tr>
<td></td>
<td></td>
<td>=0.88%</td>
<td>=0.88%</td>
<td>=0.777</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>=0.7%</td>
</tr>
</tbody>
</table>

IX. CONCLUSION

Maximum precision value with maximum f feature values generates the minimum recall value where precision is ambiguity during detecting sentence boundary, f feature values is made by ambiguous behavior of punctuation marks. This precision and f feature values are creates a minimum recall value as less sentence boundary of text, i.e. weak boundary extracts various features of text segment with less sentence boundary detection.

X. FUTURE WORK

Machine learning method of SVM must be implemented to compiler design as creating software of sentence boundary detection in all type of ambiguity of punctuation marks and symbols to English, Hindi Text, it will come into automatic sentence boundary detection sentence task and by there meaningful individual sentence are easy to recognize.

REFERENCES


