A Hybrid Approach to Identify Sentence Boundaries

Vivek Dubey

Shri Shankaracharya College of Engineering and Technology, Bhilai, Chhattishgarh, India
e-mail: vivekdubey22@ gmail.com

Abstract

Now a day, every one are using internet and enriching their knowledge. Simultaneously, accessing internet has increased Natural language Applications. Words and Sentences are important tools for Natural Language Processing. This paper explores the problem of identifying sentence boundaries and model developed for this task using hybrid approach.

Keywords: *Natural Language Processing, Word, Sentence.*

1 Introduction

THESE days Electronics text is readily available via World Wide Web (WWW) in form as Microsoft Word, Adobe PDF, Past Script, HTML, SGML, XML, etc and the preformatter are available in market to convert such documents into a plain text, however such text are noisy [1], [3], [4], [7], [10] and have to be preprocessed for Natural Language Application.

The plain text is a sequence of characters [2], [6] such as capital alphabet (A-Z), small alphabet (a-z), numeric letter (0-9) and special symbol (*, +, ?, etc) in form of heading, words, number and sentences.

Normally, Sentence is considered as an important building block in the majority of Natural Language Processing (NLP) tasks likes Parts of Speech (POS) tagging, Parsing, Document Processing, Machine Translation (MT), Text Alignment, Information Extraction (IE), Text Summarization, Smart Editor, Text to Speech conversion, etc.

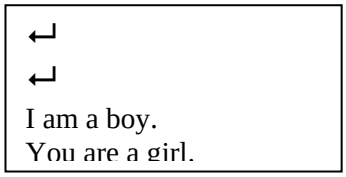
Before any syntactic analysis of the original plain text, two transformations usually take place: isolated words, called as tokenization and isolated sentence, called as sentence splitter.

There are two types of tokens formed by character structure like punctuation, number, dates, etc and other type like went, they, school, etc. The second type will undergo a morphological analysis. Sentence is a sequence of words, number and end of sentence – period, exclamation mark and a question mark. [8], [9], [11] proposed techniques to identify sentence boundary using maximum entropy, IBM word alignment model 1. In the paper, Hybrid approach, both rule based and example based, is used for identifying abbreviated token.

The rest of the paper is organized as follows. Section 2 discusses cases related sentence boundary. Section 3 introduces an overview of developed system. Section 4 presents system algorithm. Section 5 illustrates the experiment and results. Section 6 gives conclusion and further work.

2 Cases Related to Sentence Boundary

A. if there are new-lines or blank-spaces before sentences like below.



B. If a sentence are in multiple lines as like below.

I↵
am↵
a ↵
boy.

C. If after a sentence, there are new-lines or blank-spaces as like below.

I am a boy.
↵
↵
What is your name?

D. If sentences are ended with different full-stop.

I am a boy.
What is your name?
Hi!

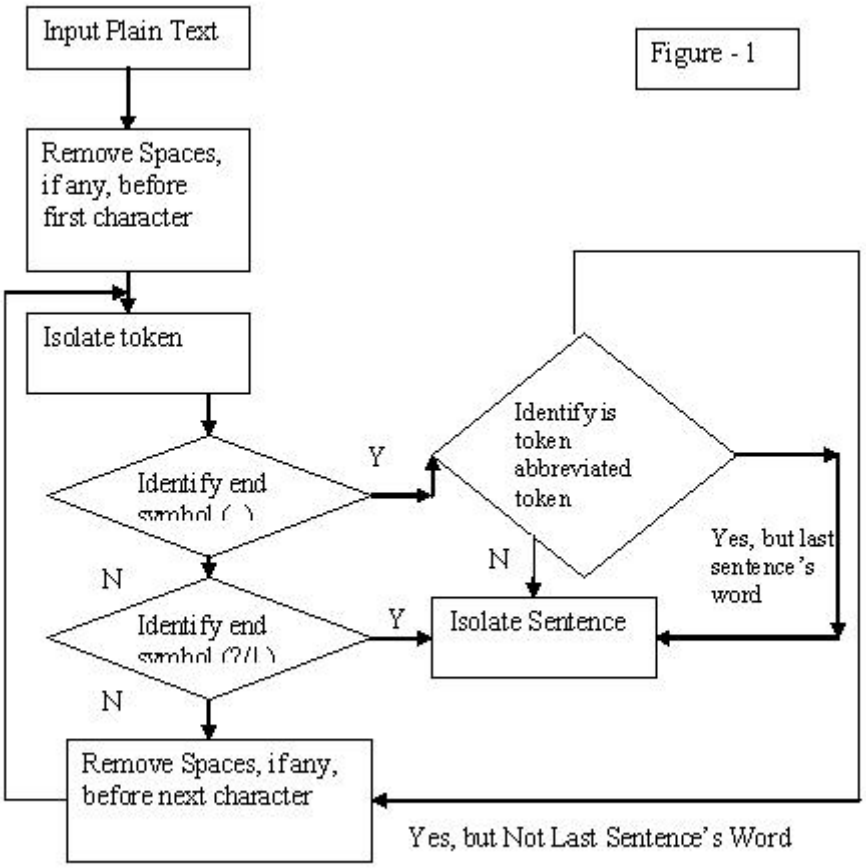
E. If sentences are with abbreviation(s).

I know **Mr.** and **Mrs.** John.
I am doing **Ph.D.**
My college name is **S.S.C.E.T.**
Are you Mr. **A.K.**Kapoor?
I will meet at **4.30 p.m.**

3 Development System Description

For identifying sentence boundary, the following system architecture was as shown in figure 1.

The process begins by introducing plain text as input to system. Then the system isolates word and if word includes symbol period (.), then identifies whether the word is abbreviated word. If word is abbreviated, then further find boundary of sentence otherwise assume boundary of sentence. During identifying sentence boundary, if Tab, blank-spaces or new-line are occurred, then system also take out such white-spaces and finally isolate sentence from given plain text.



4 Algorithm for Identifying Sentence Boundaries

While end of input file
{
Read a character
If character is not blank-space or tab or new-line
Make token
If found symbol (.)
{If abbreviated token
If last token
Isolated sentence
Else
Add token in sentence
Else
Isolated sentence
}
If found symbol ? or !
Isolated sentence
}

5 Helpful Hints Experiments and Results

The current number of abbreviation in the example database is about 210. The collection of examples from other domains is in progress. I have tested this on Brown Corpus (5) and find 50415 sentences. For this application, there are simple six steps to implement the problem as explained

below.

Step1: In following input text, there are five independent sentences.

> I know Mr. I know Mr. and Mrs. John . I am doing Ph.D. My name is R.S.Kumar. I will meet you at 7.30 p.m.

Step2: Read character as 'I' since after I there is blank-space, so make token "I" and add in sentence.

I	\0	\0	\0	\0

Step3: Read next character as 'k' since after 'k' there is next character as 'n', so add in token as "kn", similarly make "kno", "know". Since after 'w', there is blank-space, so add token in sentence as below.

I	Know	\0	\0	\0

Step4: Read next character as 'M' since after 'M' there is next character as 'r', so add in token as "Mr". Since after 'r', there is sentence end character (.), and after it, there is blank-space, so first add character (.) in token as "Mr." and then identify token from database of examples of abbreviation whether it is abbreviated token or not.

Step5: Since it is abbreviated token, so add in sentence and increment in End Symbol.

I	Know	Mr.	\0	\0

Step6: Now the next step is to find whether abbreviated token is last token of sentence or not. Read next character as "I', and assume as token, since next character is blank-space. As after identifying that token is proper now and capital letter, now assume abbreviated token "Mr." as end of sentence and add new-line as "\n" for isolate sentence from given text as shown below.

I	know	Mr.	\n	\0

Like wise, similar steps will be performed till end of input file and output file will be obtained as below.

I	know	Mr.	**\n**	I
Know	Mr.	and	Mrs.	John
.	**\n**	I	am	Doing
Ph.D.	**\n**	My	name	Is
R.	S.	Kumar.	**\n**	I
Will	meet	you	At	7.30
p.m.	**\n**			

6 Conclusion and Further work

Identifying Sentence Boundary (ISB) using Hybrid approach has been proposed. ISB matches abbreviated examples and also by rule find out abbreviated tokens, means ISB uses both methods example based and rule based as hybrid approach. The feasibility of ISB has been shown by implementing a system which isolate English sentence. The result of the experiment was encouraging. The system has been written in object oriented C++ language. As further work too, such approach can be used for other language.

References

[1] Clark, "*Pre-processing Very Noisy Text*", Proc.of Workshop on Shallow Processing of Large Corpora, 2003.

[2] G. Grefenstette and P. Tapanainen "*What is a word, what is a sentence? Problems of tokenization*", In Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94), 1994.

[3] IJCAI, "*Workshop on Analytics for Noisy Unstructured Text Data*",2007.

[4] J. Tang, H. Li, Y. Cao, and Z. Tang. " *Email data cleaning*", Proc. of SIGKDD, 2005.

[5] L. Burnard. "*Users Reference Guide for the British National Corpus*", Oxford University Computing Services, 1995.

[6] Mikheev, Andrei, "*Periods, Capitalized Words, etc.*" Computational Linguistics, 28(3), 2002 , 289–318.

[7] R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards,"*Normalization of nonstandard words*", WS'99 Final Report., 1999,

http://www.clsp.jhu.edu/ws99/projects/normal/

[8] Reynar, Jeffrey C. and Adwait Ratnaparkhi. "*A maximum entropy approach to identifying sentence boundaries*", In Proceedings of the Fifth ACL Conference on Applied Natural Language Processing, 1997, pages 16–19,Washington, D.C.

[9] Silla Jr., Carlos N. and Celso A. A. Kaestner, "*An analysis of sentence boundary detection systems for English and Portuguese documents*", In Proceedings of CICLing 2004, pages 135–141, Seoul, Korea.

[10] W. Wong, W. Liu, and M. Bennamoun, "*Enhanced Integrated Scoring for Cleaning Dirty Texts*", Proc. Of words, WS'99 Final Report., 2007

http://www.clsp.jhu.edu/ws99/projects/normal/.

[11] Xu, J., Zens, R., and Ney, H. "*Sentence Segmentation Using IBM Word Alignment Model 1*", in Proceedings of the European Association for Machine Translation, 10th Annual Conference (EAMT), 2005, Budapest, Hungary.

Article received: 2009-04-27