Journal of Engineering and Applied Sciences 13 (21): 8986-8992, 2018
ISSN: 1816-949X

Rule Based Sentence Segmentation of Indonesian Language

Suwanto Raharjo, Retantyo Wardoyo and Agfianto E. Putra
Department of Computer Science and Electronics,
Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract: Sentence detection also known as sentence boundary detection or sentence boundary disambiguation is one of the study fields in linguistic computation and one of the important stages in the development of an application or research based on natural language processing. Researches topic on Sentence Boundary Detection or Sentence Boundary Disambiguation (SBD) for Indonesian language were not get much intention by researchers as result there are not many paper are written with this topic. The Indonesian language sentence segmentation problems considered as not a big issues and could be using an English SBD method. There are could be the reasons why this topic is not get attention. Existing researches are not specially, discussed on Indonesian language sentence segmentation but only mention as one of stages of research. Two methods, rule based and machine learning are usually used as sentence segmentation methods in several languages. The other methods are using statistic based such as maximum entropy, regression tree or using artificial neural network. This study intended to do sentence segmentation using rule based method on text Indonesian language and comparing the result with existing sentence segmentation softwares. Two models of experiment are conducted on developed rules, first, using input sentences that contain ambiguity problems and second using of many sentences from several kind of input.

Key words: Sentence detection ambiguity, SBD, language Indonesian informatics, input, method, research

INTRODUCTION

Along with increasing and easily to found sources of text data, research on linguistic computation is also escalating. Text data are used in many computer science studies, one of them is linguistic computation. Sentence detection also known as Sentence Boundary Detection or Sentence Boundary Disambiguation (SBD) is one of the study fields in linguistic computation and one of the important stages in the development of an application or research based on Natural Language Processing (NLP). Several stages are based on a sentence such as Part of Speech (POS), syntactic analysis (parsing), translator machine and others. The main process of sentence segmentation performs identification of sentences in one or more paragraph.

Research on sentence segmentation methods are not many being mentioned by researchers, this situation could be happened because the high accuracy result of existing research methods or considered as not a challenging topic. A sentence detection process itself cannot be considered as a trivial task because many processes are depend on this process especially on NLP. Improvement of a sentence detection method would make direct impact to NLP result. In the future, NLP would play role especially in integrating with machines. NLP and software engineering have huge opportunity for both academicians and industries (Sharma and Yalla, 2016). Sentence segmentation process of Indonesian language is having similarity with sentence segmentation in English. Detection of sentence from one paragraph or more is done by looking on the start and end of a sentence. Sentence in Indonesian language text begins with a capital letter and ending with a mark such as period (.), exclamation (!), question (?) or quote mark. Ambiguity problems especially, the using of period mark are mostly found in SBD. Period mark could be used as the end mark of a sentence but also could be used as a part of the sentence itself.

MATERIALS AND METHODS

The common methods that are used in English sentence segmentation could break into three classes:

- Rule based
- Supervised machine learning
- Unsupervised machine learning

Some techniques in machine learning are using artificial neural network and regression tree (Palmer and

Corresponding Author: Suwanto Raharjo, Department of Computer Science and Electronics,
Universitas Gadjah Mada, Yogyakarta, Indonesia

Hearst, 1997). Statistic, maximum entropy method are also used to found the right sentence boundary (Reynar and Ratnaparkhi, 1997).

Rule based method for sentence segmentation of Indonesian language is commonly having similarity with English. Rule based pattern matching method is used in this research to do sentence detection. Several rules are built to detect a sentence by recognizing the pattern of sentence that having ambiguity problems from input text. Defined patterns that lead ambiguity problems are used to guide pattern matching processes.

The testing of rules are conducted by inputting sentences that having ambiguity problems, especially, used of period marks. The ability of rules to detect the correct pattern of end sentence are mandatory. The sentences are also tested using sentence segmentation software such as Splita, Punkt and others, then the results are compared. Text from abstract papers, online news articles and Indonesian language translation of Al-Quran are also used as data text input for testing.

Previous research: There are little amount of research references related to sentence segmentation of Indonesian language subject. Segmentation method that used to split a paragraph into sentences only mentioned at glance in a broader research topics. Sentence segmentation is done manually in the development of an Indonesian-English parallel corpus research (Larasati, 2012). Simple method to recognize sentence by detecting period, exclamation or question mark is used in Indonesian text summarization (Fachrurrozi *et al*., 2013).

Using software as API, function or online program are also used to do sentence segmentation for Indonesian text such as the using of OpenNLP (Mangasi *et al*., 2014). Another researches are not clearly mentioned the methods that are used to do sentence segmentation and the detail results of previous works are not given. Those researches do not give a detail result of sentence segmentation process are understandable because the researches are not focus on sentence segmentation process. More detail results are found on English SBD researches, success rate result of English SBD from some papers are said to have more than 98% (Reynar and Ratnaparkhi, 1997; Wong *et al*., 2014).

Sentence segmentation: Determining how sentences were build from a text for further processing is usually refer as sentence segmentation (Dale *et al*., 2000). Studied on sentence segmentation in English itself is already done for a long time especially for speech recognition. Sentence segmentation in text was mention in research by Palmer and Hearst (1997) and Reynar and Ratnaparkhi (1997). Another languages also has been research such as Thai, Chinese, Kanji and others.

Written language that using less punctuation mark or use many period mark will present challenge to do sentence segmentation. Indonesian language like English as written language having a punctuation system such as using of period, exclamation, question or quote mark as end of sentence. Indonesian as written language that having many punctuation marks present surprising problems in sentence segmentation.

The Indonesian language sentence is defined as a unit that begins with a capital letter and ends with a period, exclamation or question mark (Markhamah, 2013). Text writing in Indonesian language is guided by special rules, written into a guided book called Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan (EYD rules) (Moeliono and Dardjowidjojo, 1988). The correct writing of Indonesian language text must follow that EYD rules. The use of punctuation mark such as period, quotation, exclamation, question etc are also regulated in the guided book. Some of the rules on writing a period mark in Indonesian sentences could lead to ambiguity problems. One of the examples is the use of a period mark, the period mark could be as a part of sentence or a marker of end sentence. Another example is how to wrote the period mark after a letter of list, chart or summary and also on abbreviation or acronym that could raise ambiguity problems.

Given an example of sentence, "Mantan Wakil Menteri Agama RI, Prof. Dr. Nasaruddin Umar, Dikukuhkan Sebagai Imam Besar Masjid Istiqlal, Jakarta." That sentence is having 3 period marks, after abbreviation of Prof and Dr and behind of Jakarta word. The using of quotation marks also could raise problems for computer based sentences detection. Quotation marks or also called as quotes, quote marks, quotemarks, speech marks or talking marks are primarily used to indicate material that represents quoted or spoken language. Quotation marks also could be as the end mark of sentence if the last quote is preceded by a period, question or exclamation mark. For example, a sentence, Pasal 36 UUD 1945 berbunyi, "Bahasa Negara Ialah Bahasa Indonesia. Bahasa Indonesia adalah bahasa persatuan." Thus, the end of sentence in Indonesian text must having some requirements to be full filled.

Sentence segmentation rules: The start and end of a sentence in text would be easy to detect by human readers but would be tricky if readers are machines or computers. Computers would be able to detect sentence boundary by developing set of rules in a computer program. EYD rules could be used to guide the

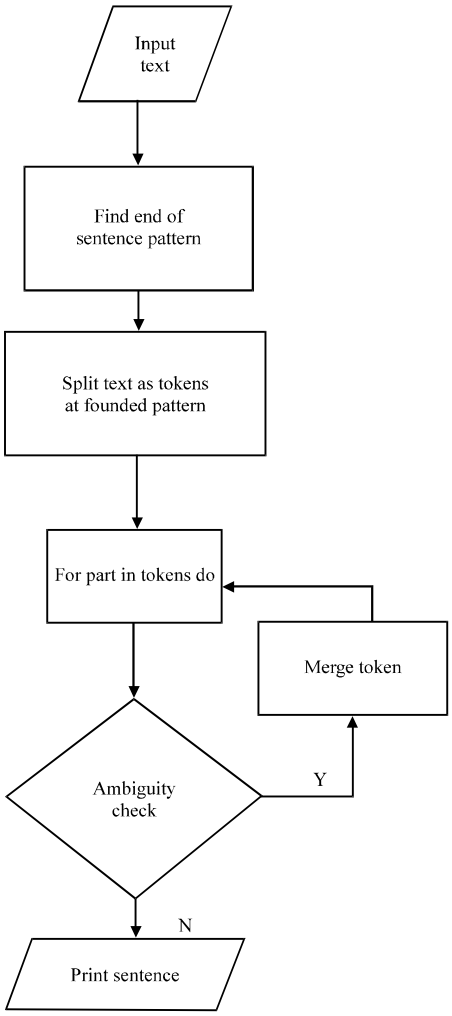

Fig. 1: Sentence segmentation process

development of algorithm for detecting sentence boundary in Indonesian language. Illustration of this research is shown on Fig. 1 which primary goal is developing of rules to do sentence segmentation of Indonesian language.

The processes shown on Fig. 1 are starting by a process that looking a pattern of end sentence from input text. Patterns of end sentence are group of characters that potentially having end of sentence marker. End marks of sentence in Indonesian language are period, exclamation and question mark. Patterns of end sentence in Indonesian language could be break into 4 type markers:

- End mark of sentence followed by space and a capital letter
- End mark of sentence followed by space, end quote mark and a capital letter

```
EndMark is one of End Sentence Marker
tokens = EndMark.split(text)
```

Fig. 2: Pseudocode find end of sentence pattern

- End mark of sentence followed by space and a capital letter
- End mark of sentence followed by end quote mark, space, start quote mark and a capital letter

Pseudo code of this process is shown at Fig. 2. Four type markers of end sentence are searched from input text, if end of pattern is match with one of markers then sentence is split into tokens. One token is text before a space and the other one is after it. Each tokens are a candidate of sentence and will be examined in the next processes for ambiguity check.

An example of text input: "Liverpool ditahan imbang 0-0 oleh West Ham United di pertandingan putaran keempat Piala FA tahun ini. Laga ini telah berlangsung di Stadion Anfield, Sabtu (30/1)." The pattern marker type (1) is founded in the example that is a period mark followed by space and a capital letter (. L). Two tokens are created on this process, first is "Liverpool ditahan imbang 0-0 atas West Ham United di pertandingan putaran keempat Piala FA tahun ini." and second one is "Laga ini telah berlangsung di Stadion Anfield, Sabtu (30/1)."

The end of sentence pattern is checked for all tokens in the example no one of tokens are contain ambiguity problems. The result is all tokens are considered as a sentence. Token or group of tokens are decided as a sentence is determined by developed rules. The main task of rules on this process is able to detect ambiguity problems on founded tokens.

Ambiguity detection: The main task of sentence segmentation is able to pick the correct of end sentence mark. Ability to choose the correct of end sentence mark could be done by solved the ambiguity problems. Common ambiguity problems are founded in Indonesian sentence text is the using of many period marks in a sentence. One of them is the using of period mark in abbreviations. Several potential ambiguity problem in abbreviations are:

- Period mark in degree abbreviation at beginning of person name
- Period mark in degree abbreviation at end of person name and last sentence
- Period mark in abbreviation of first person name
- Period mark in abbreviation of foreign term

Abbreviation of degree abbreviation or first name at beginning of person name would generate pattern of end sentence. The pattern is "period followed by space and a capital letter", this pattern would be lead to ambiguity problem. An example, given a sentence "Prof. Dr. Nasarudin Umar Adalah Imam Besar Masjid Istiqlal". The pattern of end sentence a period followed by space and a capital letter are founded in the sentence. There are a period after Prof followed by space and capital letter (D) and a period after Dr followed by space and capital letter (N). All of founded patterns are not the correct pattern for end of sentence an will lead problems in sentence detection. By developing set of rules, such of ambiguity problems can be avoided. Several rule are developed on this research to avoid ambiguity problems. The rules are:

- Rule for handling degree or titles abbreviations
- Rule for handling first or last name abbreviations
- Rule for handling numbering
- Rule for handling common abbreviation and foreign terms

Writing degree abbreviation in Indonesian text could be divided into two types, first is writing degree abbreviation in front of person name. There are not many degree abbreviations in this type, list of them are Drs., Dra., Dr., dr., Ir. and Prof. The second one is degree abbreviations wrote after last name. There are so many degree abbreviations for this second type, for example, M.Kom., S.H., S.E., S.Si., etc. The writing of those two types degree abbreviations would lead to ambiguity problems in sentence detection. An example writing a degree abbreviation and followed by first name such as "Prof. Dwi" will produce one of end sentence pattern markers that is period mark-space-capital letter.

Developed rules have to able to decide the token is part of sentence or end of sentence. If the first type degree abbreviations are founded than the rules have to decide that token is part of sentence and merge it with next token. Rule also must work if a sentence contain with more than one first type degree or title abbreviations, for example, "Prof. Dr. H. Djamaludin Adalah Psikolog Ternama". The abbreviations in the example of sentence are:

- Two degree abbreviations, that are Professor (Prof.) and Doctor (Dr)
- One Hajj title abbreviation (H)

All three abbreviations are led to ambiguity problems because contain a pattern marker for end of sentence.

```
for token in tokens:
  if LastToken.search(token)
     if AbbreDegA(token) :
       Sentence.append(token)
       ...
```

Fig. 3: Pseudocode detect degree abbreviation

The developed rules have to recognize degree or title abbreviations that heading person name and manage tokens as part of sentence not end of sentence. Figure 3 is shown pseudo code to detect degree abbreviations that wrote before first name.

The abbreviations of person's name also could lead to ambiguity problems, for example, given a sentence, "Presiden John F. Kennedy Berasal Dari Brookline". Period after a letter F is followed by capital letter K this is common pattern for end of sentence but on this example it is not end of sentence pattern. The developed rules on this research are only able to decide the token is part of sentence if abbreviation of person name is at the beginning of sentence.

The developed rules can not yet detected if person name abbreviations are found at other particular place on a sentence. Given sentence example, "Tingginya angka kematian ibu dan gizi buruk membuat Menteri Kesehatan Prof. Nila F. Moeloek memikirkan cara mengatasinya". Two sentence would be created by the rules, first sentence is "Tingginya angka kematian ibu dan gizi buruk membuat Menteri Kesehatan Prof. Nila F." and second one is "Moeloek memikirkan cara mengatasinya". The result of process is not correct because "Nila F. Moeloek" is a person name.

The developed rules failed to pick correct pattern when founded tokens are contain middle or last person name abbreviations and their positions are in the middle or last of a sentence. Having information of Part of Speech (POS) would help to resolve this problem, further research will used POS to detect the sentence boundary. Another ambiguity problems such as numbering foreign term and common abbreviations writing are handled by rules.

RESULTS AND DISCUSSION

Two models of experiment are conducted on developed rules. First, experiment using rules that are assigned to do segmentation task for 4 input sentences that contain ambiguity problems. Second, developed rules are also tested to do segmentation of many sentences from several kind of input. Four input of sentences used for testing in the first experiment are:

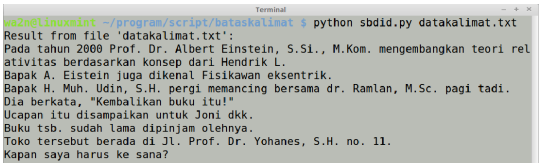

Fig. 4. Running and result of rule based program

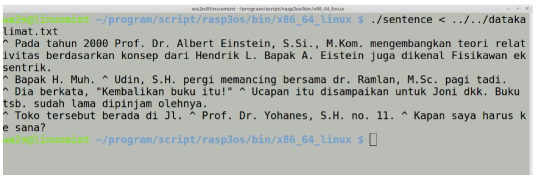

Fig. 5: Running and result of splita program

Pada tahun 2000 Prof. Dr. Albert Einstein, S.Si., M.Kom. mengembangkan teori relativitas berdasarkan konsep dari Hendrik L. Bapak A. Eistein juga dikenal Fisikawan eksentrik.Bapak H. Muh. Udin, S.H. pergi memancing bersama Dr. Ramlan, M.Sc. pagi tadi.

Dia berkata, "Kembalikan buku itu!" Ucapan itu disampaikan untuk Joni dkk. Buku tsb. sudah lama dipinjam olehnya. Toko tersebut berada di Jl. Prof. Dr. Yohanes, S.H. No. 11. Kapan saya harus ke sana?

All sentences are saved into a text file, named datakalimat.txt and newline is used as the delimiter. Beside using developed rules the sentences also being tested using SBD software and comparing the results. Four SBD software are used to test the sentences are:

- Punkt is using unsupervised machine learning (Kiss and Strunk 2006)
- Splita is using machine learning supervised
- Rasp3os is using rule based method (Briscoe *et al*., 2006)
- SpaCy is using syntactic parse tree

Python based program is used to developed rules for sentence detection. Example of execution of rule based program and the result with datakalimat.txt as input parameter is shown on Fig. 4.

Figure 4 shown all sentences are correct detected. An example of SBD ready used program, Splita when it is executed to do sentence segmentation using datakalimat.txt as input parameter is shown on Fig. 5.

A sentence progam as part of Raps3os when it is executed to do sentence segmentation using datakalimat.txt as input parameter is shown on Fig. 6.

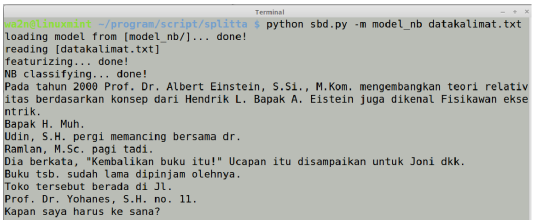

Fig. 6: Running and result of rasp3os program

Table 1: Result testing of 4 input

	Sentence detection			
Methods	(1)	(2)	(3)	(4)
Rule based	Succeed	Succeed	Succeed	Succeed
Splita (nb)	Failed	Failed	Failed	Succes
Punkt	Failed	Failed	Failed	Failed
Raps3os	Succeed	Failed	Failed	Failed
SpaCy	Failed	Failed	Failed	Failed

The others program are also executed to do sentence segmentation using same data input. The summary result of first testing using 4 input sentences are shown on Table 1.

Result on Table 1 shows almost segmentation sentence software are failed to do sentence segmentation of Indonesian language which are having ambiguity problems. As result shown, developed rules are succeed to do sentence segmentation process. All SBD ready program which used in this testing are not built for Indonesian language, it could be a reason of failed result when doing segmentation process.

Developed rules are also tested to do segmentation of many sentences from several input is the second testing. The source text input are from 150 abstract papers, online news text from Indonesian news agency (Antara) dated 4-5 February 2016 and Surah 2-3 of Indonesia translation of Al-Quran. Two treatments are applied for each data test on this second testing, first using raw data a data input without any modification. Second data test are modified according with EYD rules, especially, the writing of abbreviation in Indonesian text. Any error of the result from testing will be examined, two types possible error from this testing are:

- False positive, end of sentence mark is considered as end of sentence but actually is not
- False negative, sentence should be ended but not detected

Errors rate are calculated form how many resulted sentences divide by total of falses, the result is shown on Table 2.

Number of created sentences on segmentation process are 1884 sentences from input of 150 abstract

Table 2: Result of error detection

Input	Error rate (%)	Sentence detected	False positive	False negative
Abstract text	1.85	1184	15	7
Modified abstract text	0.25	1168	0	3
Online news articles	0.17	1155	0	2
Modified news articles	0.00	1157	0	0
Translation of Al-Quran	0.41	972	4	0
Modified translation of Al-Quran	0.00	976	0	0

papers. Sentences having pattern as follows, all sentences are begin with capital letter, 1167 sentences are ended with period, 1 sentence ended with question mark, 1 sentence ended with quote mark and 15 sentences are ended without end of sentence mark. Further examine shown that from 15 sentences are came from last sentence of abstract and ended without any end of sentence marks.

The amount of 1155 sentences are built from 103 Antara online news data input. Number of sentences that beginning with capital letter are 902 sentences and 253 started with a quote. Number of sentences ended with a quote are 20 and period as end off sentence used in 1135 sentences.

The test using Indonesian translation of Al-Quran data input, Surah No. 2 and 3 result 972 sentences. Almost sentences are begin with capital letter (971) and only one sentence is begin with a quote. Period mark used as end of sentence found on 918 sentences, quote mark is on 26 sentences, 25 sentences ended with question mark and 3 sentences using exclamation mark as end of sentence.

Table 2 shown on abstract paper data test generate 7 false negatives error, 4 sentences are forgot to inserted a space after period at end of sentence and 3 sentences contain ambiguity problems that can not yet be resolved in this rule. Fifteen false positives errors are generated from 14 false writing of period mark on abbreviation of companies name and 1 false using period mark on Latin name abbreviation. Error percentage from unmodified text data input is 1.85% or having success rate 98.15% with number of build sentences are 1184. Using of period mark after a capital letter in abbreviation in last of sentence is not yet resolved on this rules.

Modification of abstract paper, grammar writing correction especially on using of period on abbreviation adjusted according EYD rules decrease error rate to 0.25% or increase success rate to 99.75% with build sentences are 1168 sentences.

Testing of news online data input results 2 false negatives errors. One sentence begins with number that is not allowed according EYD rules, number as beginning sentences must be written as a word. Second false negative error is using of left double quotation mark that not detected by rules. False positive error is not found on testing of this online news data input. Error rate is 0.17% or success rate is 99.83% with number of build sentences are 1155 sentences. Error writing correction of this data input increase success rate to 100% with 1157 sentences made.

Third data input is Indonesian translation of Al Quran, Surah No. 2 and 3 results 972 sentences that having 4 false positive. Error came from 3 sentences that are not beginning with capital letter and 1 sentence forget to insert a space. One sentence begins with a quote and 971 are begun with a capital letter. End of sentence marks are period on 918 sentences, question mark on 25 sentences, exclamation mark used by 3 sentences and quote on 26 sentences.

Result shows that rule based method with data input using Indonesian translation of Al Quran produce error rate 0.41% or success rate is 99.59% on unmodified data input. Modification of data input increase of success rate to 100% with no false negative error.

CONCLUSION

Mostly English SBD programs are failed to do sentence segmentation of Indonesian language which are having ambiguity problems. Two SBD program success to do sentence segmentation on one of four text data input. The rule based method capable do sentence segmentation on Indonesian text that having ambiguity problems. The results also show rule based method can be relied on sentence segmentation with input data from abstract paper, online news and Indonesian translation of Al-Quran. The built method in generally is capable to do sentence segmentation in correct way, if text data input are follow the EYD rules. Results also show that the developed rule based method having some weakness. First, rule based can not do sentence segmentation correctly if data input do not follow the EYD rules. Second, rule based method only known standard ASCII quote mark (Code No. 34). Third ambiguity problems on abbreviation still became obstacle in order to do sentence segmentation especially, the using of period mark in person name and foreign term abbreviation at particular position in a sentence. Overall, more than 98% of success rate for unmodified input data text and more than 99% of success rate for inputting data text are modified according the EYD rules. Testing with more varieties text genre and size are needed to get better information.

ACKNOWLEDGEMENT

The researchers would like to acknowledge valuable funding provided by Ministry of Technology Research and Higher Education under PDD Grant 2017.

REFERENCES

Briscoe, T., J. Carroll and R. Watson, 2006. The second release of the RASP system. Proceedings of the COLING-ACL Conference on Interactive Presentation Sessions, July 17-18, 2006, ACM, Sydney, Australia, pp: 77-80.

Dale, R., H. Moisl and H. Somers, 2000. Handbook of Natural Language Processing. Marcel Dekker, New York, USA., ISBN:0-8247-9000-6, Pages: 947.

Fachrurrozi, M., N. Yusliani and R.U. Yoanita, 2013. Frequent term based text summarization for bahasa Indonesia. Proceedings of the 2013 International Conference on Innovations in Engineering and Technology (ICIET'13), December 25-26, 2013, Sriwijaya University, Bangkok, Thailand, pp: 30-32.

Kiss, T. and J. Strunk, 2006. Unsupervised multilingual sentence boundary detection. Comput. Ling., 32: 485-525.

Larasati, S.D., 2012. IDENTIC corpus: Morphologically enriched Indonesian-English parallel corpus. Proceedings of the 8th International Conference on Language Resources and Evaluation, May 21-27, 2012, Istanbul Lutfi Kýrdar International Convention and Exhibition Center, Istanbul, Turkey, ISBN:978-2-9517408-7-7, pp: 902-906.

Mangasi, T., A. Erwin and H.P. Ipung, 2014. Defined entity extraction based on Indonesian text document. Proceedings of the 2014 International Conference on ICT For Smart Society (ICISS'14), September 24-25, 2014, IEEE, Bandung, Indonesia, ISBN:978-1-4799-6323-2, pp: 61-65.

Markhamah, 2013. Ragam from Analis Kalimat Bahasa Indonesia. Muhammadiyah University Press, Indonesia,.

Moeliono, A.M. and S. Dardjowidjojo, 1988. [Standard Indonesian Language]. Ministry of Education and Culture, Indonesia, Asia,.

Palmer, D.D. and M.A. Hearst, 1997. Adaptive multilingual sentence boundary disambiguation. Comput. Ling., 23: 241-267.

Reynar, J.C. and A. Ratnaparkhi, 1997. A maximum entropy approach to identifying sentence boundaries. Proceedings of the 5th Conference on Applied Natural Language Processing, March 31-April 03, 1997, ACM, Washington, DC., USA., pp: 16-19.

Sharma, N. and P. Yalla, 2016. Software engineering and natural language processing-how can they be together?. Intl. J. Software Eng. Appl., 10: 389-396.

Wong, D.F., L.S. Chao and X. Zeng, 2014. iSentenizer: Multilingual sentence boundary detection model. Sci. World J., 2014: 1-10.