REPUBLIK INDONESIA
KEMENTERIAN HUKUM DAN HAK ASASI MANUSIA

SURAT PENCATATAN CIPTAAN

Menteri Hukum dan Hak Asasi Manusia Republik Indonesia, berdasarkan Undang-Undang Nomor 28 Tahun 2014 tentang Hak Cipta yaitu Undang-Undang tentang perlindungan ciptaan di bidang ilmu pengetahuan, seni dan sastra (tidak melindungi hak kekayaan intelektual lainnya), dengan ini menerangkan bahwa hal-hal tersebut di bawah ini telah tercatat dalam Daftar Umum Ciptaan:

I.	Nomor dan tanggal permohonan	:	EC00201705790, 24 November 2017
II.	Pencipta		
	Nama	:	**Syopiansyah Jaya Putra & Muhammad Nur Gunawan**
	Alamat	:	Jalan Pulo Mas II H No. 6 RT. 005/012 Jakarta 13210, Jakarta Timur, DKI JAKARTA, 13210
	Kewarganegaraan	:	Indonesia
III.	Pemegang Hak Cipta		
	Nama	:	**Syopiansyah Jaya Putra & Muhammad Nur Gunawan**
	Alamat	:	Jalan Pulo Mas II H No. 6 RT. 005/012 Jakarta 13210, Jakarta Timur, DKI JAKARTA, 13210
	Kewarganegaraan	:	Indonesia
IV.	Jenis Ciptaan	:	Karya Tulis (Artikel)
V.	Judul Ciptaan	:	**Sentence Boundary Disambiguation for Indonesian Language**
VI.	Tanggal dan tempat diumumkan untuk pertama kali di wilayah Indonesia atau di luar wilayah Indonesia	:	20 November 2017, di Salzburg
VII.	Jangka waktu perlindungan	:	Berlaku selama hidup Pencipta dan terus berlangsung selama 70 (tujuh puluh) tahun setelah Pencipta meninggal dunia, terhitung mulai tanggal 1 Januari tahun berikutnya.
VIII.	Nomor pencatatan	:	05219

Pencatatan Ciptaan atau produk Hak Terkait dalam Daftar Umum Ciptaan bukan merupakan pengesahan atas isi, arti, maksud, atau bentuk dari Ciptaan atau produk Hak Terkait yang dicatat. Menteri tidak bertanggung jawab atas isi, arti, maksud, atau bentuk dari Ciptaan atau produk Hak Terkait yang terdaftar. (Pasal 72 dan Penjelasan Pasal 72 Undang-undang Nomor 28 Tahun 2014 Tentang Hak Cipta)

a.n. MENTERI HUKUM DAN HAK ASASI MANUSIA
REPUBLIK INDONESIA
DIREKTUR JENDERAL KEKAYAAN INTELEKTUAL
u.b.
DIREKTUR HAK CIPTA DAN DESAIN INDUSTRI

Dr. Dra. Erni Widhyastari, Apt., M.Si.
NIP. 196003181991032001

Sentence Boundary Disambiguation for Indonesian Language

Syopiansyah Jaya Putra[1], Muhamad Nur Gunawan[2], Ismail Khalil [3], Teddy Mantoro [4]
[1,2]Faculty of Science and Technology, Syarif Hidayatullah State Islamic University Jakarta
[3]Institute of Telecooperation, Johannes Kepler University Linz,
[4]Faculty of Engineering and Technology, Sampoerna University
syopian@uinjkt.ac.id[1], nur.gunawan@uinjkt.ac.id[2], ismail.khalil@jku.at[3], teddy@ieee.org[4]

ABSTRACT
Sentence boundary detection is essential for natural language processing (NLP). Sentence boundary detection in the Indonesian language has lots of problems, which includes punctuation, abbreviation, and character in the bracket. The disambiguation should be detected as sentence boundary. Thus the sentence boundary system can divide the sentences accurately. This study presents the development of a training dataset for the existing model to optimize supervised sentence boundary detection for the Indonesian language. Indonesian Translation of the Quran (ITQ) data set was used in this study by using the supervised method. The following is the process briefly: create the training data, apply sentence detection to separate sentences on ITQ, and calculate precision, recall, and F-measure. The result is quite promising, it gives as follows: Precision of 91.7%, Recall 81.6%, and F-Measure 86.4 %, respectively.

CCS Concepts
- **Information systems → Information retrieval → Retrieval tasks and goals → Information extraction**

Keywords
Sentence boundary detection; natural language processing; Indonesian text; sentence disambiguation, supervised optimization.

1. INTRODUCTION
Sentence boundary detection is very important for natural language processing (NLP). The text processing used by NLP depends on the Sentence Boundary Detection in solving the sentence disambiguation problem. Sentence Boundary Detection in Indonesian has many problems including punctuation, abbreviation, and characters in the bracket. There are available a number of NLP tools for English, however in applying to the Indonesian language, it's not working as expected.

Sentence boundary disambiguation considers as a task of identifying the sentence elements within a paragraph or an article. Sentence boundary disambiguation is one of the essential problems for many applications of NLP, including parsing, information extraction, machine translation, and document summarizations [1].

The sentence has an essential structure in the text of documents, and this structure is vital for NLP [2-4]. The meaning or knowledge from text document can be defined based on the structure of sentences. More studies have been done in the area of sentence boundary detection for natural language processing [4]. Those include sentence boundary detection for the text of Canada [1], Vietnamese [5], Croatian [6], Polish [7], Chinese [8, 9] and English [10, 11]. However, there are no studies available yet for sentence boundary detection in Indonesian language.

Sentence boundary detection system has three methods [4, 10], i.e., rule-based, unsupervised, and supervised method. Rule-based model [4] has several approaches, include Stanford parser [12], LingPipe, Bora's Model, and Mikheev's Model. Unsupervised methods used Hidden Markov Model (HMM) [3, 13] and Hidden Topic Markov Model (HTMM) [14]. The supervised method may also utilize Riley's model and Maximum Entropy (ME) model [5,11,15-17].

Supervised methods which proposed by Riley [6] presents a system for disambiguating sentence endings that end with a period using a decision tree classifier. In 1997, Reynar and Ratnaparkhi showed a solution based on a Maximum Entropy (ME) model for the sentence boundary detection (SBD) problem. The main advantages of their approach are convincing by a mathematical approach, unfortunately, only a little essential information for disambiguation works. The ME model that they proposed did not require POS tags, heuristic rules, or domain-specific information, such as the list of abbreviations and proper names. Instead, the system used many diverse features extracted from the local context. The model can attain an accuracy of 98.8% on the Wall Street Journal data, which is quite powerful; consider how simple the model is and how flexible it can select the features.

The study on sentence boundary detection on supervised methods based on Maximum Entropy can be divided into three parts [18]: (1) MaxEnt classifier, (2) feature extraction, and (3) data sets. MaxEnt classifier can be used to predict sentence boundaries, MaxEnt produced a probability distribution over the possible labels based on features extracted from local contexts in the training data. These probabilities were computed using an objective function and its derivatives, which are smoothed to prevent over fitting. The test labels then predicted using the features extracted from their contexts and the corresponding probability distribution over the labels. Feature extraction can be used trigrams context that occurred within the training sentence. The data sets have three common characteristics, i.e., a list of sentences, the words in each sentence, and their corresponding part of speech tags.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
iiWAS '17, December 4–6, 2017, Salzburg, Austria
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5299-4/17/12…$15.00
https://doi.org/10.1145/3151759.3156474

As general knowledge, a sentence is a sequence of words ending with a terminal punctuation, such as a '.', '?', or '!'. Most sentences use a period in the end. However, sometimes a period can be associated with an abbreviation, such as "*dr.*" or represent a decimal point in a number like 10.30 (at ten thirty) [19]. That kind of punctuation, abbreviation, and number is a problem for sentence boundary detection.

The data sets the Indonesian language were examined from the Indonesian Translation of Al-Qur'an (ITQ), which contains a collection of documents in the Indonesian language that are structurally different from English.

In ITQ the sentence variation consists of short sentences, interrogative sentence, long sentences, and a group of sentences. First, a short sentence is sentencing with subject – object and punctuation "." at an end, e.g., *Raja Manusia.* (The King of mankind). Second, the interrogative sentence is sentencing with predicate-question – subject with punctuation "?" as an end, e.g., *Dalam keadaan bagaimana kamu ini?* (In what condition were you?). Third, long sentence is a sentence with subject – predicate – object – time/place and punctuation "." at an end, e.g., *dan kaum Tsamud yang memotong batu-batu besar di lembah.* (And with Thamud who carved rocks in the valley). Forth, a group of sentencing is sentences with subject-predicate-object and punctuation "," as separated and "." as an end, e.g. *Hanya Engkaulah yang kami sembah, dan hanya kepada Engkaulah kami meminta pertolongan* (It is You we worship and You we ask for help).

Sentence boundary detection in the Indonesian language is very important because of lots of studies, such as text processing, novelty detection, clustering, categorization, question-answer system, use the Indonesian document as their object study. Indonesian alphabet can be seen as it is not different from the English alphabet, also the boundary of sentences that is similar to the English text, such as a period. So, this study will use supervised sentence boundary detection with some additional calculation to fit the Indonesian documents.

This study proposes an optimizing supervised sentence boundary detection model for analyzing Indonesian sentences boundary disambiguation. Since supervised method still has a high F-Measure [20] especially using public tools such as OpenNLP, this study creates a training data in Indonesian language to optimize the accuracy of sentence boundary detection and tested using OpenNLP.

The publicly open tools OpenNLP [4] used the supervised method to find sentence boundary detection. This approach can be divided into three step: (1) create the training data, (2) apply sentence detection on OpenNLP to separate sentences on ITQ, (3) use the evaluator to calculate precision, recall, and F-Measure.

At the end of this paper, this study evaluates the result by discussing two scenarios. First, it used default train data on OpenNLP, and second, it has optimization processed the train data in the Indonesian language.

2. RESEARCH METHOD

2.1 Data sets

For this study, 114 Indonesian text documents were used from Indonesian Translation Quran (ITQ). These data set were collected from http://tanzil.net/#trans/id.indonesian with the last update on June 4, 2010. ITQ is a significant religious text translated from Quranic Arabic into the Indonesian language. ITQ contains 6,236 numbered verses (*ayat*) and is divided into 114 chapters.

2.2 Model description

The supervised method to find sentence boundary detection is divided into three step. First, the data set was used to make the training data. Second, the sentence boundary detection algorithm was used those training data. Third, the result of sentence boundary detection was evaluated by using Sentence Boundary Evaluator.

Step 1. In this step, since the train data by default using English sentence, the training data set was built to fit in the Indonesian language model. From this point, the Sentence Boundary Annotation [19] was considered to be used in determining the potential *sentence boundary symbols* (SBS). For Indonesian translation language texts, the sentence boundary symbols (".", "!","?","?") and bracket symbols (")","}") can be used as sentence boundary. In specific documents, the symbol ("|") was used as sentence boundary as well.

The punctuation and symbols can be combined and blended into training data file. Thus, the training data can be compiled in OpenNLP format. Therefore, the file of binary features of Indonesian sentence, i.e. itq-sent_trained.bin, as a file to be used for next step.

Step 2. After the training data was produced, OpenNLP was used by the following process:

1. Load binary file (itq-sent_trained.bin);
2. Input document (ITQ);
3. Run the OpenNLP model;
4. Run sentence detector;
5. Get the result;

In this step, OpenNLP model uses binary features of Indonesian sentence to detect the sentence boundary detection in ITQ document file.

Step 3. For evaluation, the sentence boundary detection, Precision, Recall and F-Measure were calculated for both supervised method on OpenNLP and modified train data respectively.

The experimental testing was measured and then evaluate the performance such as in [4]. Where Precision is the fraction of the retrieved documents which are relevant. Recall is the fraction of the relevant documents which have been retrieved, tp (true positives) is a number of relevant elements which have been retrieved, fp (false positive) are a number of irrelevant element which has been retrieved, and fn (false negative) is a number or relevant elements which have not been retrieved.

The F-Measure values are within the interval [0, 1] and larger values indicate higher detection quality. By these measures, overall precision and recall values as well as an overall F-Measure value, were computed as the average mean of the precision, recall and F-Measure values for all documents.

3. RESULT AND DISCUSSION

This development has been detected 6236 lines of text of the document. The sentences result, divide into two parts. First part is sentenced boundary detection with default train data on

OpenNLP. Second part is sentence detection boundary detection using modified train data which contain the Indonesian language.

For the first experiment, the input document (id.indonesia.txt) giving the result such as the following:

77|46|(Dikatakan kepada orang-orang kafir): "Makanlah dan bersenang-senanglah kamu (di dunia dalam waktu) yang pendek; sesungguhnya kamu adalah orang-orang yang berdosa".

((It is said to the disbelievers): ”Eat and enjoy yourselves (in the world) is short; indeed, you are sinners”

For the sentences above, uncorrected sentence boundary detection was found, such as the sentence after ":" and ";" is not separated. From the whole result, the problems are similar with the three above. So, by using default training data in OpenNLP, 6230 sentences were not correct and only 3318 sentences were correct as the result of training data using Indonesian sentences.

For the second experiment, the input document (id.indonesia.txt) giving the result such as the following:

1. *Dikatakan kepada orang-orang kafir*

   (It is said to the disbelievers)

2. *Makanlah dan bersenang-senanglah kamu (di dunia dalam waktu) yang pendek*

   (Eat and enjoy yourselves (in the world) is short)

3. *sesungguhnya kamu adalah orang-orang yang berdosa*

   (Indeed, you are sinners)

For the three sentences above, corrected sentence boundary detection was found. As for Sentence Boundary Detection by using training data in the Indonesian language gives a significant result, it has Precision: 91.7%, Recall: 81.6%, and F-Measure: 86.4%. This result is important that the document produce more than 80% good sentences. It means that the training data for detect Sentence Boundary Detection in the Indonesian language show the reliable result.

By adding appropriate sentences to the data train, the sentence boundary detection can produce appropriate sentences.

The ability of sentence boundary detection in English has achieved an accuracy of 98.8% [15]. Therefore sentence boundary detection in the Indonesian language requires an improvement in the future.

4. CONCLUSION

This paper presents a solution to sentence boundary detection problem in Indonesian Language, which includes punctuation, abbreviation, and character in the bracket. This study described the development of a training dataset for the existing model to optimize supervised sentence boundary detection in the Indonesian language. The disambiguation has been shown in the discussion that it capable to detect sentence boundaries by dividing into several sentences/phase with relevant boundary. ITQ data set has been used by using the supervised learning method by the following process: create the training data, apply sentence detection to separate sentences on ITQ, and calculate precision, recall, and F-measure. In addition to the data set, the abbreviation is also included in the data train.

By using valid ITQ train data, this study capable to detect and generate sentence boundary accurately.

In the future, the train data can be used for sentence boundary detection for other Indonesian documents.

5. REFERENCES

[1] M. Parakh, N. Rajesha, and M. Ramya, "Sentence Boundary Disambiguation in Kannada Texts," *Language in India, www. languageinindia. com, Special Volume: Problems of Parsing in Indian Languages*, pp. 17-19, 2011.

[2] N. Wanjari, G. Dhopavkar, and N. B. Zungre, "Sentence Boundary Detection for Marathi Language," *Procedia Computer Science*, vol. 78, pp. 550-555, 2016.

[3] B. Jurish and K.-M. Würzner, "Word and Sentence Tokenization with Hidden Markov Models," *JLCL*, vol. 28, pp. 61-83, 2013.

[4] J. Read, R. Dridan, S. Oepen, and L. J. Solberg, "Sentence boundary detection: A long solved problem?," *COLING (Posters)*, vol. 12, pp. 985-994, 2012.

[5] H. P. Le and T. V. Ho, "A maximum entropy approach to sentence boundary detection of Vietnamese texts," in *IEEE International Conference on Research, Innovation and Vision for the Future-RIVF 2008*, 2008.

[6] F. Šarić, J. Šnajder, and B. Dalbelo Bašić, "Optimizing Sentence Boundary Detection for Croatian," in *Text, Speech and Dialogue*, 2012, pp. 105-111.

[7] P. P. Mazur, "Text segmentation in Polish," in Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on, 2005, pp. 43-48.

[8] Y. Yang and N. Xue, "Chinese comma disambiguation for discourse analysis," in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 2012, pp. 786-794.

[9] N. Xue and Y. Yang, "Chinese sentence segmentation as comma classification," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2, 2011, pp. 631-635.

[10] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Computational Linguistics*, vol. 32, pp. 485-525, 2006.

[11] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the fifth conference on Applied natural language processing*, 1997, pp. 16-19.

[12] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenic, "Triplet extraction from sentences," in *Proceedings of the 10th International Multiconference" Information Society-IS*, 2007, pp. 8-12.

[13] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, 2002, pp. 1-8.

[14] A. Gruber, Y. Weiss, and M. Rosen-Zvi, "Hidden topic Markov models," in *Artificial intelligence and statistics*, 2007, pp. 163-170.

[15] A. Ratnaparkhi, "A simple introduction to maximum entropy models for natural language processing," *IRCS Technical Reports Series*, p. 81, 1997.

[16] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *INTERSPEECH*, 2002.

[17] A. Ratnaparkhi, "Maximum entropy models for natural language ambiguity resolution," *A Dissertation, University of Pennsylvania*, 1998.

[18] N. Agarwal, K. H. Ford, and M. Shneider, "Sentence boundary detection using a maxEnt classifier," in *Proceedings of MISC*, 2005, pp. 1-6.

[19] K. Tomanek, J. Wermter, and U. Hahn, "Sentence and token splitting based on conditional random fields," in *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, 2007, pp. 49-57.

[20] D. J. Walker, D. E. Clements, M. Darwin, and J. W. Amtrup, "Sentence boundary detection: A comparison of paradigms for improving MT quality," in *Proceedings of the MT Summit VIII*, 2001.