RESEARCH ARTICLE

Advanced Science Letters
Vol. 23, 12186–12189, 2017

Text Processing Application for Indonesian Documents

Putu Manik Prihatini* and I. Ketut Suryawan

Electrical Engineering Department, Politeknik Negeri Bali, Jl. Kampus Bukit Jimbaran, Kuta Selatan, Badung, 80364, Bali, Indonesia

The processing text of document is a complex process and plays an important role for the next process in information retrieval system. It requires a complete knowledge base and appropriate grammar structures. The knowledge base and grammar cannot be generally accepted for all kind of language. Unfortunately, some researches on text processing in Indonesian language until now have not been able to generate a knowledge base as a reference for research. Even, stemming algorithms that was developed still have limitations in defining the rules of the morphology. Therefore, through this research, was developed the text processing application for Indonesian documents using Matlab software. The application consists of tokenization, filtering, stemming and evaluation. The application also generates two knowledge bases such as stop list and dictionary of word base. The result of stemming in this application was tested using metric of Precision and Recall, and has resulted the value at 0.97 and 0.64. The result of this research is expected to be used in the next process of text mining, and also as a reference in research on the information retrieval system for Indonesian documents.

Keywords: Text Processing, Knowledge Base, Stemming Algorithm, Indonesian Documents.

1. INTRODUCTION

Information has grown very rapidly, both of quantity and quality because information is needed by the human. The information can be delivered not only in hardcopy, but can be acquired digitally via the Internet. Search engine is used by humans for searching digital information. Search engine has many collection of documents, and also has a text box that used by user for typing query. The documents and query are processed for looking the similarities of information. Based on this similarity, documents are grouped into several clusters. Last, the answer is determined by ranking the documents in the best clusters.

Search engine process is very complex because it requires a complete knowledge base and appropriate grammar structures. The knowledge base and grammar cannot be generally accepted for all kind of language.

This has led to the information produced by search engine cannot satisfy users.[1] The information that are displayed by search engine was not really capable to direct the user to the appropriate document. This is because of knowledge base that used by search engine is not complete.[2] Furthermore, the parsing of grammar rules on the stemming process has an important role in finding the root of each word in the document.[3–8]

Some of researches have been generating text processing application for Indonesian language, but it still have limitation of knowledge base.[9–11] Currently, there is no Indonesian knowledge base that can be used as a reference for research such as WordNet for English.[12, 13] Some of researches have been generating algorithm of stemming process for many language.[14–20] Research on stemming algorithms have been done for Indonesian language, but many of those have not been published, so it is difficult used as reference for research.[10, 21–25] Furthermore, the evaluation result of measurement metric for the Indonesian stemming algorithm is still not optimal. Therefore, it needs more researches to generate text processing application to produce a complete knowledge base and stemming algorithm for Indonesian language.

Through this research, it will generate text processing application for Indonesian documents, including tokenization, filtering, stemming and evaluation. The result of this paper can be used by other researchers for the next process on text mining like feature extraction. The result is also expected to be a reference in studying information retrieval system for Indonesian documents.

2. METHODOLOGY

Text processing application architecture for Indonesian documents can be described in the research design as shown in

*Author to whom correspondence should be addressed.

 1936-6612/2017/23/12186/004 doi:10.1166/asl.2017.10598

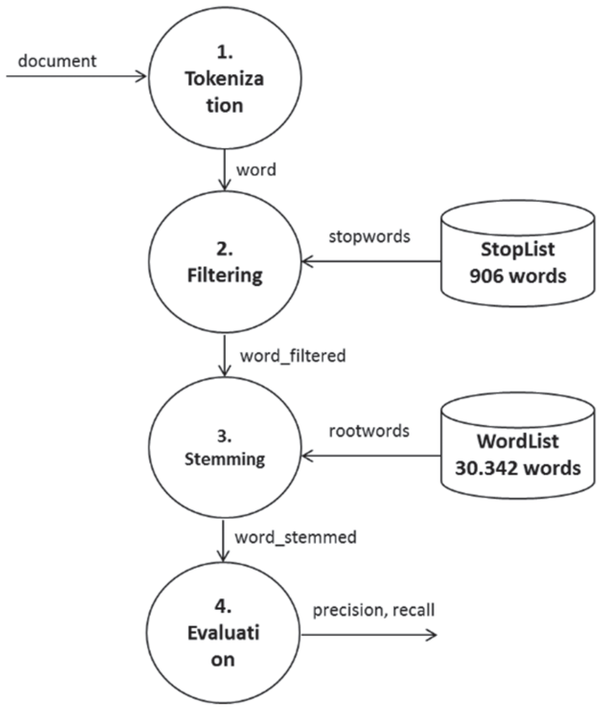

Fig. 1. Text processing application architecture.

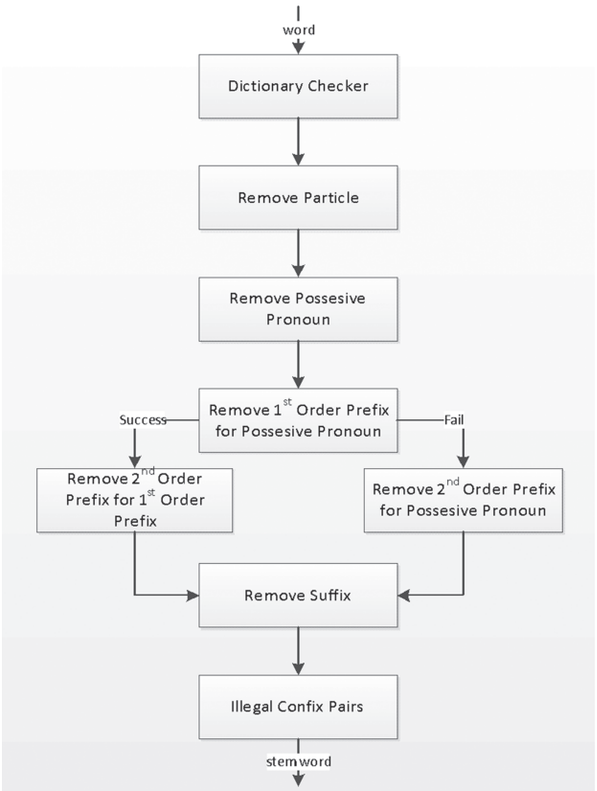

Fig. 2. Stemming process.

Figure 1. The architecture is designed consists of four processes such as tokenization, filtering, stemming and evaluation.

2.1. Tokenization

Tokenization is a process to decompose the text of a document into sentences and decomposing its back into words. Tokenization begins with changing the text into lowercase. Then, tokenization performed by removing number, punctuation and space. This process produces a set of terms (or words) from document.

2.2. Filtering

Filtering is the process to remove words that have no meaning to the content of the document. The words that have no meaning is known as stop word and stored in stop list. In this research, stop list contain stop word and the most common words that published by Tala.[24] Considering the test data used comes from online media, so that the addition of stop list made to words that often appear in the article that has no meaning to article content. Number of words used in the stop list are 906 words. Some examples of the added word from the test data such as 'aja (just),' 'akrab (familiar),' 'apa–apa (anything),' 'biar (let it),' 'sih (hmm),' 'gitu (so),' and others.

2.3. Stemming

Stemming is the process of finding the root of a word in a document. For Indonesian language, stemming has done by removing prefixes and suffixes. Therefore, this process requires a dictionary of word base as a reference for removal of prefixes and suffixes. In this research, the number of word used in dictionary is 30.342 words.

The rules of removal in this research using the morphological structure for Indonesian language that published by Tala.[24]

$$[\text{prefix1}][\text{prefix2}] + \text{root word} + [\text{suffix}][\text{possessive}][\text{particle}]$$

To process these rules, it takes a stemming algorithm. This research use Porter-like Stemmer for Indonesian algorithm that published by Tala.[24] But, there is a modification has been done for the algorithm. At the beginning of process, the checking of each word in dictionary has been added to algorithm, as shown in Figure 2.

Based on morphological structure above, it requires knowledge of the rules for prefix and suffix in Indonesian language. This research use five rules which combines the rules that published by Tala[24] and M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams.[23] Table I show the deletion rules for the suffix including particle pronouns, possessive pronouns and suffix. Table II shows the deletion rules for the first and second order of derivational prefix, with its replacement and additional condition. Table III shows a pair of affixes that are not allowed in Indonesian grammar.

Table I. Inflectional and derivational suffixes.

Inflectional particle	Inflectional possesive	Derivational suffixes
-kah	-ku	-kan
-lah	-mu	-an
-pun	-nya	-i

Table II. The first and second order of derivational prefix.

First order	Replacement	Followed by	Second order	Replacement	Followed by
meng	null	null	ber	null	null
meny	s	vowel	bel	null	'ajar'
men	null	null	be	null	k+er
mem	p	vowel	per	null	null
mem	null	null	pel	null	'ajar'
me	null	null	pe	null	null
peng	null	null			
peny	s	vowel			
pen	null	null			
pem	p	vowel			
di	null	null			
ter	null	null			
ke	null	null			
se	null	null			

Table III. Illegal confix pairs.

Prefix	Suffix
ber	i
di	i|an
ke	i|kan
meng	an
peng	i|kan
ter	an
be	i
me	an
se	i|kan
te	an

2.4. Evaluation

In this research, the evaluation of text processing application performed on the results of stemming. This evaluation using metric measurements of Precision and Recall.[26] Precision indicates the degree of accuracy among the information requested by the user and the answers given by the system, with the equation as shown in (1). Recall indicates the success rate of the system in rediscovering information, with the equation as shown in (2). TP is a true positive, FP is a false positive, and FN is a false negative.

$$p = \frac{TP}{TP + FP} \tag{1}$$

$$r = \frac{TP}{TP + FN} \tag{2}$$

3. EXPERIMENTAL RESULT

3.1. Data Preparation

Application testing is performed on the stemming process using 125 news data obtained through digital media Detikcom. The example of article is shown in Figure 3. The news is manually saved into a text file (*.txt). The files are grouped into five categories, where each category is divided into five sub-categories, as shown in Table IV.

3.2. Implementation

Text processing application in this research was built using Matlab software. The evaluation process is done by comparing the results of stemming generated by the system and the results of stemming obtained manually.

Cari Berita atau Alamat

Foto: Mendag saat memantau harga daging di pasar (Ardan/Detik)

Jakarta -Pemerintah tengah terus berupaya agar harga bahan pokok dapat turun dan stabil menjelang Lebaran. Namun, ada komoditas seperti daging sapi segar yang masih tinggi di angka Rp 120.000/kg.

Padahal, Presiden Joko Widodo (Jokowi) meminta harga daging sapi bisa dijual Rp 80.000/kg selama Ramadan dan Lebaran.

Menteri Perdagangan (Mendag), Thomas Lembong (Tom Lembong) mengakui saat ini pemerintah masih belum optimal mengimpor daging sapi sehingga harga daging sapi masih di atas Rp 100.000/kg.

"Memang pasok yang kami impor masih belum mencukupi. Ini hanya untuk mulai mengguyur pasar, mulai meredam pasar. Tentunya kita tidak bisa hanya berhenti di sini. Lanjut terus," ungkapnya usai acara pencanangan program pengampunan pajak oleh Presiden RI di Kantor Pusat Direktorat Jenderal Pajak, Jakarta, Jumat (01/07/16).

Saat ini pemerintah melakukan kebijakan dengan menambah kuota impor daging sapi. Kebijakan ini diharapkan dapat menurunkan dan menstabilkan harga.

"Dalam minggu-minggu berikutnya jumlah pasok yang akan didatangkan lebih meningkat dari puluhan ton menjadi ratusan ton. Dan dalam sebulan dua bulan akan menjadi ribuan ton. Jadi baru kita bisa harapkan

Fig. 3. News article digital on detikcom.

Precision and recall value in this research indicates the average of precision and recall for 125 files. This precision and recall value also compared with the result of stemming that has been done by Tala[24] and Adriani et al.,[23] as shown in Table V.

Precision value at 0.96779629 (97%) shows the accuracy degree between the results obtained by the system and the results obtained manually. Recall value at 0.642687871 (64%) shows the success rate of system in stemming a word.

3.3. Analysis

The average of precision and recall value was good enough, but still not optimal. This is due to several reasons, such as:

1. The number of words used in dictionary of word base and stopwords used in stoplist is still inadequate
2. Morphological structures in the Indonesian language has ambiguity
3. There is no knowledge base for objects such as the person's name, city name, company name, name of the day, etc.
4. Human errors when typing a word in the original document
5. There are too many foreign words (that are not in Indonesian language), numbers and terms in the original document.

Table IV. Categorization of data.

Category	Sub category
Finance	Industry, Finance Services, Monetary Policy, Entrepreneur, Property
Entertainments	Fashion and Style, Movie, Music, International Artist, National Artist
News	Criminal, Government, Defense and Security, Politic, Social and Cultural
Sport	Weight-Lifting, Bike-Racing, Badminton, MotoGP, Football
Technologies	Photography, Gadget, Games, Social Media, OS and Software

Table V. Comparison of results.

Algorithm	Precision	Recall
Adriani et.al.	0.7026	0.6563
Tala	0.7086	0.6574
This research	0.9678	0.6427

4. CONCLUSIONS

Text processing application for Indonesian documents in this research consist of tokenization, filtering, stemming and evaluation. This research also produces two knowledge bases, such as stoplist and dictionary of word base. The application was built using Matlab software. The evaluation process has resulted average value of precision at 97% and average value of recall at 64%. This shows that the system is good enough to stem the word, but it needs to be improved again to increase the recall value. The result of this research is expected to be used in the next process of text mining, and also as a reference in research on the information retrieval system for Indonesian documents.

For the future work, the research will modify stemming algorithm used in this research, so that the problem of ambiguity in the morphology rules of Indonesian language can be handled.

Acknowledgments: This research was supported by grants of DIPA Politeknik Negeri Bali Indonesia SP Dipa-042.01.2.401006/2016.

References and Notes

1. Y. Chali, S. A. Hasan, and M. Mojahid, *Information Processing and Management* 51, 252 **(2015)**.
2. F. Ahmad, M. Yusoff, and T. M. T. Sembok, *Journal of the American Society of Information Science* 47, 909 **(1996)**.
3. P. Willett, *Program* 40, 219 **(2006)**.
4. C. D. Paice, Stemming, Encyclopedia of Language and Linguistics, Second edn. **(2006)**, pp. 149–150.
5. M. A. Fattah, F. Ren, and S. Kuroiwa, *Information Processing and Management* 42, 1003 **(2006)**.
6. B. Fox and C. J. Fox, *Information Processing and Management* 38, 547 **(2002)**.
7. A. Pirkola, *Journal of Documentation* 57 **(2001)**.
8. W. Kraaij and R. E. Pohlmann, *The New Review of Document and Text Management* 1, 25 **(1995)**.
9. W. Suwarningsih, I. Supriana, and A. Purwarianti, *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 13, 357 **(2015)**.
10. S. D. Larasati, V. Kubon, and D. Zeman, Indonesian morphology tool (MorphInd): Towards an Indonesian corpus, *Proceedings of Systems and Frameworks for Computational Morphology-Second International Workshop, SFCM 2011*, Zurich, Switzerland **(2011)**.
11. A. Purwarianti, M. Tsuchiya, and S. Nakagawa, *IEICE Transactions on Information and Systems* E90-D, 1841 **(2007)**.
12. G. A. Miller, *Communications of the ACM* 38, 39 **(1995)**.
13. G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, *International Journal of Lexicography* 3, 235 **(1990)**.
14. F. N. Flores and V. P. Moreira, *Information Processing and Management* 52, 840 **(2016)**.
15. A. Arslan, *Information Processing and Management* 52, 326 **(2016)**.
16. T. Brychcín and M. Konopík, *Information Processing and Management* 51, 68 **(2015)**.
17. B. Abuata and A. Al-Omari, *Journal of King Saud University-Computer and Information Sciences* 27, 104 **(2015)**.
18. L. Dolamic and J. Savoy, *Information Processing and Management* 45, 714 **(2009)**.
19. N. L. Bhamidipati and S. K. Pal, *EEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics* 37, 350 **(2007)**.
20. M. Bacchin, N. Ferro, and M. Melucci, *Information Processing and Management* 41, 121 **(2005)**.
21. M. Widjaja and S. Hansun, *International Journal of Technology* 6, 139 **(2015)**.
22. D. Suhartono, D. Christiandy, and R. Rolando, *Journal of Software* 9 **(2014)**.
23. M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, *ACM Transactions on Asian Language Information Processing (TALIP)* 6, 1 **(2007)**.
24. F. Z. Tala, A study of stemming effects on information retrieval in Bahasa Indonesia, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands **(2003)**.
25. V. B. Vega and S. Bressan, Indexing the Indonesian web: Language identification and miscellaneous issues, *Tenth International World Wide Web Conference*, Hongkong **(2001)**.
26. C. Goutte and E. Gaussier, A probabilistic interpretation of precision, recall and *f*-score, with implication for evaluation, *Proceedings of the European Colloquium on IR Research (ECIR'05)*, LLNCS 3408, Springer **(2005)**.

Received: 30 August 2016. Accepted: 30 May 2017.