WORD SENSE DISAMBIGUATION USING LESK METHOD FOR CHILDREN'S SHORT STORIES IN INDONESIAN LANGUAGE

Dina Ramandha Putri[1], Ken Kinanti Purnamasari[2]

[1,2] Informatics Engineering – Indonesian Computer University
Dipatiukur Street 112-114 Bandung
Email : dinaramandhaa@gmail.com[1], guinev_gs@yahoo.co.nz[2]

ABSTRACT

Word Sense Disambiguation is a process of identifying the meaning of the word is used in a context when the word has several different meanings. In previous research, Word Sense Disambiguation using Simplified Lesk have lower time complexity and the resulting accuracy greater than Original Lesk [1].

Word Sense Disambiguation system stage starting from the receipt of the text input documents to do the preprocessing consists of sentence tokenizing, words tokenizing and stopword removal.The result of preprocessing in the form of sentences that are processed by specifying an ambiguous word, and searching for the meaning of an ambiguous word using kateglo dictionary. If a word has more than one meaning and the kind words more noun or verb to be selected as an ambiguous word. Next, counted how many times the occurrence of the same word in the sentence and word on every ambiguous word meaning using Simplified Lesk. The large number of occurrences of the same word will produce the largest weights and a candidate selected the proper meaning to the ambiguous words.

Based on the test results using the data as much as 35 Indonesian short stories online taken from several sources, it was concluded that the Word Sense Disambiguation using Simplified Lesk produces a very small performance with an average value of 30% recall, precision of 18.82%, f-measure of 19.68%, and 14% of accuracy. It is caused due to a complete lack of said inventory in the kateglo dictionary so that many dictionary words which are not detected.

Keywords— Word Sense Disambiguation, ambiguous word , kateglo, weights, preprocessing, Simplified Lesk, Original Lesk.

1. INRODUCTION

Language is a means of communication to convey a message. The message contained in the specific meaning of a message is delivered. In a message often found ambiguity. The ambiguity that arises in the context of language, good word, phrase, clause, sentence, or a discourse. In a short story contained a sentence that has a meaning and understanding. Therefore, the interpretation of the story can arise against ambiguity. To overcome this it needs word sense disambiguition in order to identify the meaning of words in a certain context that has a number of different meaning [2].

In this study, Word Sense Disambiguation system stage starting from the receipt of the text input documents to do the preprocessing consists of sentences tokenizing, words tokenizing and stopword removal. The result of preprocessing in the form of a sentence are processed by searching for the meaning of each word with kateglo dictionary.

Words that have more than one meaning and type more than one word will be selected as the word is ambiguous. Article selanjutya counted how many times the occurrence of the same word in the sentence and word on every ambiguous ambiguous meaning of the word by using Simplified Lesk. Said that more comes up lame or overlap will produce the largest score and a candidate for the right meaning for the word ambiguous [3].

As for some of the limitations of Word Sense Disambiguation system built, including:

In this research, the word called ambiguous when the word has a word class noun (noun) or verb (verb) and the amount of meaning more than one.

a. Does not detect derived words
b. the meaning of the Election carried out based on the occurrence of the word (weights) Simplified the greatest Lesk
c. If no occurrences of the same word (weighting = 0) then the meaning of the election carried out on first order
d. If the value of the occurrence of the word of the same value in a list of meaning then done checking synonyms in the ambiguous word
e. Checking synonym do twice

Description of the system of Word Sense Disambiguation can be seen in Figure 1.

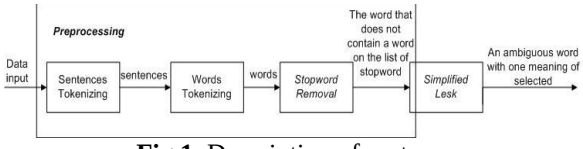


Fig 1. Description of system

2. RESEARCH CONTENT

2.1. Review of Literature

2.1.1. Word Sense Disambiguation

Word Sense Disambiguation or meaning of the word disambiguation is a process of identifying the meaning of the words used in certain sentence when the word has several different meanings.

2.1.2. Online Dictionary

Online dictionary is a website that provides a service to translate words/sentences like dictionary imaginable at the bookstore. in this site we can change a word in one language to the other. some dictionaries also give a sense of the service or the meaning/intent of a Word. Online language dictionaries currently Indonesia consists of several types such as; KBBI Online Kemdikbud, the great Indonesian Language Dictionary – the University of Indonesia, kateglo, etc.

In this study an online dictionary that is used is kateglo dictionary because the program contains quite a lot of words and meaning and sqlnya database available that facilitates the retrieval of data.

2.1.3. Kateglo

Kateglo is an application and web service source and open content for a dictionary, a thesaurus and glossary languages of Indonesia. Its name is taken from the acronym element service: ka (mus), te (saurus) and glo (sarium) [7].

2.1.4. Preprocessing

Stage of processing is to prepare input data which will be processed on the next step. Research on preprocessing consists of several stages, namely: sentence tokenizing, words tokenizing and stopword removal. The process of preprocessing can be seen in Figure 2.

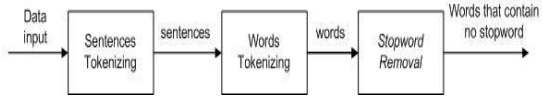


Fig 2. Preprocessing

Sentence tokenzing is used to separate the text into the form of a unit of the sentence. So that each sentence will serve as a measure for the determination and calculation of the ambiguous wording on Simplified methods Lesk. Tokenisasi used to separate words from each sentence be unit said. Stopword removal is the process of the Elimination of common words that are not to be used and processed in the Simplified Lesk.

2.1.5. Performance Testing

The testing that was done in this study using performance testing. Standard measurements are commonly used in the study of Word Sense Disambiguation i.e. recall, precision, and f-measure [5].

Recall is the success rate of the summary. Calculation of recall can be seen in the equation as follows:

$$recall = \frac{tp}{tp + fn} \quad \dots(2)$$

where

recall : success rate

tp : the sum of sentences that are managed in accordance with the system extracts the human extracted sentences

fn : the sum of sentences extracted human but not contained in the extracted sentences system

Precision is the degree of accuracy of the results summary. Calculation of precision can be seen in the following equation:

$$precision = \frac{tp}{tp + fp} \quad \dots(3)$$

where

precision: correctness rate

tp : the sum of sentences that are managed in accordance with the system extracts the human extracted sentences

fp : the number of sentences extracted system but not in the sentence extracted human

F-measure is a combination between recall and precision. The f-measure calculation can be seen in the following equation:

$$F - measure = \frac{2\ x\ recall\ x\ precision}{recall + precision} \quad \dots(4)$$

$$Accuracy = \frac{tp + tn}{tp + fp + fn + tn} \dots(5)$$

2.2. Data Input

Data input needed to run the system formerly included in advance into the database were sourced from Kumpulan Cerita Anak Karya Mahasiswa Pendidikan Guru Sekolah Dasar Universitas Negeri Yogyakarta [5], kidnesia.com, 22 Hari bercerita Buku Kumpulan Cerita Anak .

So the user selects the title of the short story which will be processed. If the user wants to choose the story itself can be done with the data added to the data is just .txt file. One example of the data entered to study Word Sense Disambiguation can be seen in Table 1 with the title of "Helping Neighbors".

Table 1. Data Input

Title: Helping Neighbours
Pak Hasan memiliki ritme kehidupan dan kebiasaan layaknya orang desa. Ia mempunyai rumah, sawah, kebun, dan binatang piaraannya. Binatang piaraan berupa ayam, angsa, itik, kerbau, dan sapi. Kata Pak Hasan, ayam, angsa, dan itik bisa bertelur dan mampu menghasilkan daging untuk dimakan. Telur itu bisa dikonsumsi untuk lauk makan. Sapi dan kerbau bisa untuk alat membajak sawah. Membajak bisa menghasilkan uang. Oleh karena itu, Pak Hasan merawat dan memelihara semua binatang piaraannya itu dan selalu memandikannya dengan sebuah bak tak terpakai berisi air di belakang rumahnya seperti yang dilakukan Ali, tetangganya.

2.3. *Preprocessing*

In the preprocessing phase will be done sentences tokenizing, words tokenizing and stopword removal.

1. Sentence Tokenizing

Sentence tokenizing is the process break up the text into sentences with delimeter point. So that each sentence will serve as a measure for the determination and calculation of the ambiguous wording on Simplified methods Lesk. Tokenisasi sentence on this system using the delimeter point ("."). The process of tokenisasi of the sentence can be seen in Figure 3. And the results can be seen in Tabel 3.

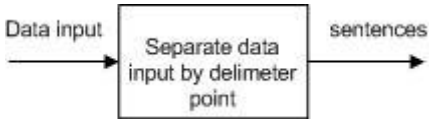


Fig 3. Sentences Tokenizing

Table 2. Sentences Tokenizing

Before	After	
Pak Hasan memiliki ritme kehidupan dan kebiasaan layaknya orang desa. Ia mempunyai rumah, sawah, kebun, dan binatang piaraannya. Binatang piaraan berupa ayam, angsa, itik, kerbau, dan sapi Kata Pak Hasan, ayam, angsa, dan itik bisa bertelur dan mampu menghasilkan daging untuk dimakan. Telur itu bisa dikonsumsi untuk lauk makan. Sapi dan kerbau bisa untuk alat membajak sawah. Membajak bisa menghasilkan uang. Oleh karena itu, Pak Hasan merawat dan memelihara semua binatang piaraannya itu dan selalu memandikannya dengan sebuah bak tak terpakai berisi air dibelakang rumahnya seperti yang dilakukan Ali, tetangganya.	**Number**	**Sentences**
	1	Pak Hasan memiliki ritme kehidupan dan kebiasaan layaknya orang desa
	2	Ia mempunyai rumah, sawah, kebun, dan binatang piaraannya
	...	...
	8	Oleh karena itu, Pak Hasan merawat dan memelihara semua binatang piaraannya itu dan selalu memandikannya dengan sebuah bak tak terpakai berisi air dibelakang rumahnya seperti yang dilakukan Ali, tetangganya

2. Word Tokenizing

Word Tokenizing is the process split the sentence into the units of the word with a space delimeter. The separation sentence using delimeter space (' '). Tokenisasi the word done to make the word into an array to ease in the search for meaning in a dictionary program. Tokenisasi process can be seen in Figure 4. And the results can be seen in TABLE III

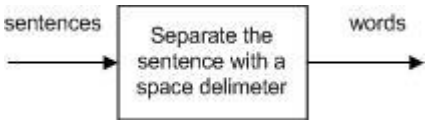


Fig 4. Word Tokenizing

Table 2. Words Tokenizing

No	Before	After	
1	Pak Hasan memiliki ritme kehidupan dan kebiasaan layaknya orang desa	Pak	dan
		Hasan	kebiasaan
		memiliki	layaknya
		ritme	orang
		kehidupan	desa
2	Ia mempunyai rumah, sawah, kebun, dan binatang piaraannya	Ia	kebun
		mempunyai	dan
		rumah	binatang
		sawah	piarannya

3. *Stopword Removal*

Stopword Removal is the process of the Elimination of common words that are not to be used and processed at later stages. Words in bold will be discarded when the word is found in the list of stopword.Stopword removal process can be seen in Figure 5.

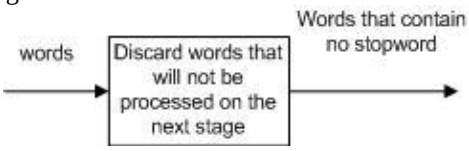


Fig 5. Stopword Removal

Table 3. *Removal Stopword*

No	Before		After	
1	**Pak**	**Dan**	memiliki	orang
	Hasan	Kebiasaan	ritme	desa
	memiliki	Layaknya	kehidupan	
	ritme	Orang	kebiasaan	
	kehidupan	Desa	layaknya	
2	**Ia**	Kebun	mempunyai	binatang
	mempunyai	**Dan**	rumah	piarannya
	rumah	Binatang	sawah	
	sawah	Piarannya	kebun	

2.4. *Simplified Lesk*

Simplified Lesk is a method to identify the meaning of ambiguous words in a context with the use of machine readable dictionary or dictionary [4].

Simplifed Lesk have connectedness will return the number of words an overlapping (score) between sentence meaning and where the biggest score will be decisive in the selection of the proper meaning. .

Calculation simplified lesk can be seen on the following formula: [9]

$$ScoreLeskvar(S) = | context(w) \cap gloss(S) | \quad (1)$$

Where:
ScoreLeskvar(S) : Score of Simplified Lesk

context (w) : Sentence which contains ambiguous words

gloss (S) : List the meaning of an ambiguous word

The first step is checking every word in a sentence use the kateglo dictionary. Basic words that have the word class noun (noun) or verb (verb), has more than one meaning and is not a derived terms are categorized as an ambiguous word. Aside from that the word is not ambiguous. The sentence already contains no words in the list of stopword.

Table 5. Determination Of Ambigous Words

Sentences	Words	Words Class	Sum of meanings
memiliki ritme kehidupan kebiasaan layaknya orang desa	Memiliki	*verb* (derrived word)	2
	ritme	*noun*	1
	kehidupan	*noun*	1
	kebiasaan	*noun*	1
	layaknya	*adverb*	2
	orang	*noun*	12
	desa	*noun*	4

Next, check the meaning of each word said that ambiguous predetermined.

Table 6. Checking the Meaning of Words

Sentences	Ambigous Words	Meanings
memiliki ritme kehidupan kebiasaan layaknya orang desa	Orang	manusia (dalam arti khusus)
		manusia (ganti diri ketiga yang tidak tentu)
		dirinya sendiri; manusianya sendiri
		kata penggolong untuk manusia
		anak buah (bawahan)
		rakyat (dari suatu negara);
		warga negara
		manusia yang berasal dari atau tinggal di suatu daerah (desa, kota, negara, dsb)
		suku bangsa
		manusia lain
		bukan diri sendiri
		bukan kaum (golongan, kerabat) sendiri
		(cak) karena (sebenarnya)

Check every word on the meaning of each word in the sentence. When in sentences and words have the same meaning or have overlapping words, count how many times the occurrence of the word that is the same on both. On the contrary, if in the sentence and the meaning of the same Word does not exist or there is no overlapping of words then the system will choose the meaning on the first order.

Table 7. Calculation of the Occurrence of the Word

Sentences	Ambigous Words	Meanings	Score
memiliki ritme kehidupan kebiasaan layaknya orang **desa**	Orang	manusia (dalam arti khusus)	0
		manusia (ganti diri ketiga yang tidak tentu)	0
		dirinya sendiri; manusianya sendiri	0
		kata penggolong untuk manusia	0
		anak buah (bawahan)	0
		rakyat (dari suatu negara);	0
		warga negara	0
		manusia yang berasal dari atau tinggal di suatu daerah (**desa**, kota, negara, dsb)	1
		suku bangsa	0
		manusia lain	0
		bukan diri sendiri	0
		bukan kaum (golongan, kerabat) sendiri	0
		(cak) karena (sebenarnya)	0

Specify a value of (the most abundant occurrence of the word). The value of the greatest occurrence of the Word will be selected as a candidate for potential meaning becoming meaning elected. If the value of the appearance of the same word (weights) on the list of the meanings of the same value then do some checking synonym to add

value to the occurrence of the word (weights) of an ambiguous word twice checking.

Table 8. Checking of Synonyms

Sentences	Ambigous Words	Meanings	Score
tiba-tiba matanya melihat buah anggur merah bergantungan	anggur	(v) menganggur	0
		(n) keratan cabang atau batang yang akan ditanam; setek	0
		(v) tumbuhan memanjat menjalar yang buahnya kecil-kecil sebesar kelereng dan berangkai	0
		(v) buah anggur	1
		(v) minuman dari sari buah anggur yang difermentasi	1

After getting the initial process like do a synonym that is check out each word on the meaning of each word in the sentence. Do stage Simplified Lesk as in Table 7**.**

Table 9. Synonyms

Ambigous Words	Synonym
anggur	Berpangku tanggan-
	Ongkang-ongkang
	Tidak bekerja

When the words do not have meaning becomes the synonym of the candidate selected and the meaning is the meaning to the second order. The greatest value (the most abundant occurrence of the word). The value of the greatest occurrence of the Word will be selected as a candidate for potential meaning becoming meaning elected.

2.5. Result

From the results of steps with Simplified Lesk method on Table 5 until Table 7 to be known Word that identified the exact meaning of the ambiguous and elected to the ambiguous words. The result can be seen in Table 10.

Tabel 10. Result

Sentences	Ambigous Words	Meanings
memiliki ritme kehidupan kebiasaan layaknya orang desa	orang	manusia yang berasal dari atau tinggal di suatu daerah (desa, kota, negara, dsb)
Telur bisa dikonsumsi lauk **dimakan**	telur	berbagai-bagai benda yang bentuknya (rupanya, sifatnya, dsb) menyerupai **telur**;
tiba-tiba matanya melihat buah anggur merah bergantungan	anggur	minuman dari sari buah anggur yang difermentasi

2.6. Testing

The testing that was done in this study using performance testing. Standard measurements are commonly used in the study of Word Sense Disambiguation i.e. recall, precision, and f-measure.

Conducted testing on 35 Indonesian Short Story of The Children[8].

Table 11. Testing

Cerpen ke-	Sum of sentences	*R*	*P*	*F*	A
1	8 sentences	40%	33.33%	36.36 %	30%
2	12 sentence	14.29%	16.67%	15.38%	15.38 %
3	12 sentence	25%	16.67%	20%	11.11 %
4	12 sentence	28.57%	33.33%	30.77%	25%
5	15 sentence	14.29%	12.5%	13.33%	17.14 %
6	14 sentence	14.29%	14.29%	14.29%	14.29 %
7	12 sentence	20%	14.29%	16.67%	9.1%
8	14 sentence	20%	22.22%	21.5%	21.5%
9	17 sentence	28.57%	33.33%	30.77%	25%
10	11 sentence	12.5%	12.5%	12.5%	6.67%

11	27 sentence	12.5%	12.5%	12.5%	6.67%
12	17 sentence	20%	16.67%	18.18%	18.18 %
13	22 sentence	16.67%	14.29%	15.38%	8.33%
14	19 sentence	10%	11.11%	10.53%	15%
15	23 sentence	25%	22.22%	23.53%	23.53 %
16	11 sentence	16.67%	16.67%	16.67%	9.1%
17	19 sentence	12.5%	12.5%	12.5%	12.5%
18	7 sentence	14.49%	12.5%	13.33%	7.14%
19	19 sentence	33.33%	33.33%	33.33%	20%
20	20 sentence	8.33%	8.33%	8.33%	4.35%
21	16 sentence	25%	25%	25%	14.29 %
22	27 sentence	16.67%	16.67%	16.67%	9.1%
23	11 sentence	22.22%	22.22%	22.22%	12.5%
24	23 sentence	20%	25%	22.22%	30%
25	26 sentence	12.5%	14.29%	13.33%	13.33 %
26	28 sentence	8.33%	7.14%	7.69%	4%
27	36 sentence	16.67%	16.67%	16.67%	9.1%
28	23 sentence	42.86%	42.86%	42.86%	27.27 %
29	15 sentence	33.33%	20%	25%	14.29 %
30	15 sentence	16.67%	16.67%	16.67%	9.1%
31	27 sentence	16.67%	11.11%	13.33%	7.14%
32	19 sentence	50%	35.71%	41.67%	26.32 %
33	17 sentence	9.1%	9.1%	9.1%	4.76%
34	17 sentence	46.15%	42.86%	44.44%	28.57 %
35	21 sentence	50%	50%	50%	33.33 %
Average		22.08%	18.82%	19.68%	14.04 %

Where
R : Recall
P : Precision
F : F-Measure
A : Accuracy

2.7. Result of Testing

From table XI Testing for Word Sense Disambiguation using Simplified Lesk produce average of recall 30%, precision 18.82%, f-measure19.68%, and 14% accuracy.

3. CLOSING

3.1. Conclusions

Based on the results of research that has been done, it can be known that Word Sense Disambiguation using Simplified Lesk methods in Indonesian Short Story of The Children has very small performances with an average of 22% recall, precision of 18.9%, f-measure amounting to 12.2%, and accuracy of 14%. The following is an explanation of the cause of the performance results of Word Sense Disambiguation that is so small that the resulting system, including:

a. Not enough many words in inventory of kateglo dictionary so that many words and meanings that are not detected.
b. The use of the synonyms that are used to add weights Simplified Lesk did not produce significant results.
c. System does not detect the noun so it's not ambiguous words he knew might also originate from derivative word.
d. An error in preprocessing stopword removal where the words in the stopword list contains ambiguous words so that word is not processed in the Simplified Lesk.

1.

3.2. Suggestions

In this research by using the kateglo dictionary as a resource in determining an ambiguous word with one meaning, not enough help increase accuracy of Word Sense Disambiguation. Further research so that it is advisable to replace the resource dictionary with the dictionary more and complete inventory list of meanings and their words have meaning, so that it can improve the accuracy of Word Sense Disambiguation.

THANKS TO

Big thanks to Miss Ken Kinanti Purnamasari who have a lot of spare time to provide guidance, advice, motivation, advice, and science during the process of the preparation of this research.

BIBLIOGRAPHY

[1] D. Craggs .J, "An Analysis and Comparison Predominant Word Sense Disambiguation Algorithm," Computer Science Honours Dissertation, Faculty of Computing Health and Science, Edith Cowan University, Western, Australia, 2011.

[2] J. Harmoejanto, "Pembentukan graf untuk pelabelan makna kata berdasarkan synset, relasi synset dan gloss", *Seminar Sistem Informasi (SESINDO2010)*, December 2010.

[3] J. Ekedahl and G. Koraljka, "Word Sense Disambiguation UsingWordNet and the Lesk Algorithm", Unpublished.

[4] A. Ranjan and D. Saha, “Word Sense Disambiguation: A SURVEY”, *J. IJCTM*. Vol. 5, No.3, July 2015.
[5] A. Com. Press, “Kumpulan Cerita Anak Mahasiswa Pendidikan Guru Sekolah Dasar (PGSD) Universitas Negeri Yogyakarta”, Wired 11 October 2016, [Online]. Tersedia:http://staff.uny.ac.id/sites/default/files/lain-lain/dr-ennyzubaidahmpd/Penyunting%20Editor%20Reviewer%20Resensi_Cerita_Anak_Petualangan.pdf [Accessed 11, october 2016]
[6] R. Navigli, “Word Sense Disambiguation: A Survey. ACM Computing Surveys”, *J. ACM Computing Survey*, Vol.41, No.2, February 2009.
[7] Anonim, “Kateglo”, Wired 20 October 2016, [Online]. Tersedia: https://kateglo.com [Diakses 20 oktober 2016]