ERRORS BY AUTO-MORPHOLOGICAL ANALYSIS IN A CHILDREN STORY CORPUS: AN EVALUATION OF MORPHIND PROGRAM

Noveka Erviana Nur Alfiani

Project Advisor : Prihantoro

Email : novekaveka@gmail.com

English Department, Faculty of Humanity, Diponegoro University
Jalan Prof. Sudarto, S.H., Tembalang Semarang 50275, Telepon: (024)

76480619, Fax: (024) 7463144, Web: www.fib.undip.ac.id

ABSTRACT

Indonesian Morphological Tool, Morphind, is meant to make a proper morphological analysis before doing further automatic language processing. Morphind is applied to enrich raw Indonesian text with morphological information, the preprocessing stage of an Indonesian corpus. In this study, the data is obtained from children's stories in the website ceritaanak.org by taking 500 types of total 2101 types. The purpose of this study is to identify and classify the types of errors present in data processing using morphind program. In the analalysis I uses the method Introspective and Dictionary Indonesian (KBBI) to validate the analysis. The findings of this research suggest that there are still many aspects that can be improved about morphind. Recommendations are fixing the data base especially for OOV (out of vocabulary) and dictionary accuracy, improving the display for the Allomorph, and improving the algorithm for morpheme segmentation.

Keywords : *Morphology, Morphind, Automatic morphological analysis, error analysis*

I. INTRODUCTION

The Indonesian language, is the official language of Indonesia. Language technology research in this language is quite encouraging lately but without a well-developed long-term plan. There are many language tools such as parsers,

semantic analyzers and speech recognition tools. The Indonesian Morphological Tool, Morphind, is meant to make a proper morphological analysis before doing further language processing. Morphind is applied to enrich raw Indonesian text with morphological information, the preprocessing stage of an Indonesian corpus.

purpose of research: to use morphind program created by computer scientists to analyze words in the data and To identify and classify the types of errors contained in data processing using the morphind program. The result of this analysis will help reader to know more about morphind program created by computer scientists to analyze words and identify the errors. I hope there will be other research an evaluation morphind program that can be used as the object of other research.

II. THEORETICAL FRAMEWORK

2.1. Morphology

Morphology is a study about word structure (Fromkin, Rodman, & Hyams, 2009). Morphology is a study that systematically learns about the internal structure of words (Haspelmath & Sims, 2010). Words can still be broken down into several more complex parts such as roots, affixes, stems and bases. Morphemes, In the science of morphology are used to identify smaller part of words. In this study I analyze the word using morphind program (Larasati, 2011) and identify the types of morphologycal errors. The word categories form-class words and structure-class words. In general, the form classes provide the primary lexical content; the structure classes explain the grammatical or structural relationship. Class classification of open class there are noun, verb, adjective and adverb, while class classification is closed there are determiner, pronoun, auxiliary, conjunction (or conjunct), interrogative, preposition, and particle.

2.2. Affix

Morphemes are the smallest grammatical units that have meaning. Free morpheme is a morpheme that can stand alone as a word, such as *tour* and *walk* (Fromkin, Rodman, & Hyams, 2009). Bound morpheme is a morpheme that can

not stand alone. Suffixes and Prefixes are examples of bound morpheme. (Plag, 2002). Affix is a bound morpheme that attaches to bases (Plag, 2002, page. 100). Lexeme is a word in the abstract sense, lexemes are abstract entities that do not have their own phonological form. While, the word form is a word in a concrete sense (Haspelmath & Sims, 2010, page. 15).

2.3. Allomorph

Allomorph is variant form of a morpheme but it does not change the meaning. Allomorph has different in pronunciation and spelling according to their condition (Plag, 2002, page. 124). It means that allomorph will have different sound, pronunciation or spelling in different condition. The condition depends on the element that it attaches to.

2.4. Token and Type

Token is simply defined as running word, while type is the distinct tokens. In type, same token are only counted one. However, when counting tokens they are counted depending on the occurences not the variation. The exmple of token and type, Mary goes to Edinburgh next week and she intends going to Washington next month. The same word of the sntence are distinct tokens of a single types. The term word would be ambiguous between a ‘type’ interpretation and a ‘token’ interpretation, the ambiguity would be just the same as is exhibited by many other terms not specifically related to language.

III. RESEARCH METHOD

In this study I use morphind program to analyze the data. In the collecting data, I extracted the types through the unitex program. Of the total data, the authors took as much as 20% from 2101 types, which is 500 types. The data is then analyzed using morphind program. The research data is categorized thematically based on the type of error.

The author used the Introspective method and Dictionary Indonesian (KBBI) reference to validate the analysis of morphind program. Introspective method is a method of providing data by utilizing the language intuition of researchers who examine the language dikuasainya (mother tongue) to provide the

necessary data for the analysis in accordance with the purpose of this research. this method is intended as an attempt to reveal the identity of the formation of language form that can allow people to carefully determine certain lingual units whose unclear-lingual status is unclear.

IV. RESULT AND DISCUSSION

In this section, the writer will show the data that has been collected and processed by Morphind program. In morphology word can be divided into several smaller components (Chaer, 2008, p. 13). There are 3 different categories of errors from Morphind. They are tagset, allomorph, and morpheme break. These three errors will be described in more detail.

4.1. Tagset

Tagset shows part of speech tags or can be called as POS tag. It is a set of writing symbols used to show POS (Larasati, 2011). Tagset errors in this corpus are divided into 4, i.e. Clitic tagset, Word or surfaceform tagset, entry tagset, and tagset that is not in the data base (OOV).

4.1.1. Clitic Tagset

Clitic is a free morpheme, where the underlying form is orthographically attached to another component or a word (Chaer, 2008, p. 5). *ku-* is one proclitic embedded in front of the base. And then, the example of an enclitic attached to the back of the word are *-nya, -mu*, Here is an example of clitic analysis that is not correct:

(1) ^lilin<n>_NSD+dia<p>_PS3$

The example show *-nya* clitic analysis by Morphind program, while examples (5), (7), (9) shows their sentences contexts. Morphind analyzed *-nya* as a third person pronoun. In example (4) the word *lilinnya, -nya* is not a pronoun. *-nya* refers to a particular candle rather than in general. The following examples are word analysis by using a new tag for the definite article, because in the

analysis of morphind program there is no tag of definite article. The author uses tag <e> for definite article. They are correct analysis of clitic :

(2) ^lilin<n>_NSD+nya<e>_ES3$

4.1.2. Entry Tagset

The basic form is a base that can take morphological process to be a word (Chaer, 2008, p. 21). In this case the error is in giving tag to the base of a full form, which that call entry tag. Entry tag is written in lowercase, such as <n>, <v>, <p>. Examples of incorrect tags for the analysis are shown in example:

(3) ^bisa<n>_NSD$

The word *bisa* is analyzed as <n> by Morphind program. However the correct tag is <v>, because in (18) *bisa* is a modal verb that describes its subject to be able to perform an activity. The correct example is:

(4) ^bisa<v>_VSD$

4.1.3. Word of Surface form Tagset

Word or surface form may take the same form like the base, or different form (undergo morphological process). In Morphind word tag is indicated by uppercase letters on the right side, which is different from entry tag, which is indicated by lowercase letters. The incorrect example is:

(5) ^bahu <v> _VSA$

The word *bahu* is a noun. But on the results of the analysis by Morphind program shows the word *bahu* is a type of verb (VSA). The tag of the word *bahu* in (22) is ideally (NPS) not (VSA). The correct example is:

(6) ^bahu<n>_NSA$

4.1.4. Tagset that are not in the Data Base

In the analysis, this type of error analysis occurs both in entry or surface form tags. The symbol of the error is with <x>, which means unknown or undefined. Consider example:

(7) ^betah<x>_X--$

The tag <x> and X are given because Morphind cannot detect the word in the data base. *betah* is supposed to be an adjective.

4.2. Allomorph

Allomorph is a term used in the field of linguistics for the variation of a form of morpheme (Chaer, 2008, p. 15). Therefore, allomorph is a realization of a real or existing morpheme. The example is:

(8) ^ber+gegas<v>_VSA$

On analyzed number (8) with the Morphind program the word *bergegas,* is wrong in writing the morpheme prefix. The morpheme is an allomorph where the writing of the R must be capitalized. The correct samples of the analysis is:

(9) ^beR+gegas<v>_VSA$

The example in (9) is example that prefix writing at the beginning of sentence is written correctly. The prefix is an allomorphic form so that in its writing, the allomorph must be written using capital letters i.e. beR-.

4.3. Morpheme Break

Morpheme break is a kind of mistake in delimiting words and the overall error of the word where the word should be written separately. This morpheme break has several categorizations or combinations of affix, suffix, and confix.

4.3.1. Morpheme Break Surface form or Boundaries

This section will analyze the error of a basic word that includes verbs, properties, objects and others as a whole where the word has been affixed either affix or suffix. There is example incorrect:

(10)^berkacamata<x>_X--$

The words in the example is affixed. There Fore they are supposed to be delimited properly. The analysis by Morphind program is still wrong. Therefore, the author wrote re-analysis of the three words such as shown in example:

(11) ^ber+kacamata<n>_V--$

kacamata that originally standalone without a prefix called a noun, in which the word is attached with prefix beR- that turn the word into *berkacamata* into a kind of verb, which *beR*- means to use. *berkacamata* becomes a verb that has the meaning of using glasses.

V. CONCLUSION

The results of research on data processing using morphind program shows that the error rate is 39% of the total data in the analysis of 500 words. There are 3 types of errors from our analysis, which can proportionally described as: 64% Tagset, 19% allomorph, and 17% morpheme break errors. Tagset error can be subdivided 4 types: Clitic, Entry, POS, OOV. Among the errors, the largest error is tagset, while the smallest number of errors occur in morpheme break.

Based on the results of the research, the researcher suggests that there are still many aspects that can be improved about morphind. Recommendations are fixing the data base especially for OOV (out of vocabulary) and dictionary accuracy, improving the display for the Allomorph, and improving the algorithm for morpheme segmentation. Chart 1 and 2 in the appendix are diagrammatic views of the error rate by the Morphind program.

REFERENCES

CeritaAnak.org. (2015, Desember 14). Retrieved Desember 14, 2015, from Ceritaanak.org.

Afini, U. (2016). Penerapan Analisis Morfologi Untuk Penanganan kata berimbuhan pada pos tagger Bahasa Indonesia Berbasis Statistik. Thesis, 72.

Chaer, A. (2008). Morfologi Bahasa Indonesia. Jakarta: PT Rineka Cipta.

Fromkin, V., Rodman, R., & Hyams, N. (2009). An Introduction to Language.

New York: Michael Rosenberg.

Haspelmath, M., & Sims, A. D. (2010). Understanding Morphology. London: An Hacette UK Company.

Larasati, S. D. (2011, August). Larasati, S. D., Indonesian morphology tool (morphind): Towards an indonesian corpus. International Workshop on Systems and Frameworks for Computational Morphology (pp. 119-129). Berlin: Springer Heidelberg.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics. New York: Cambridge University Press.

Paumier, S. (2003). Unitex 3.0 User Manual. Paris: Université Paris-Est Marne-la-Vallée.

Plag, I. (2002). Word-formation in English. Cambridge: Cambridge University Press.

Rashel, F., Luthfi, A., Dinakaramani, A., & Manurung, R. (2014). Building an Indonesian Rule-Based Part-of-Speech Tagger. 4.