Teaching Machines to Recognize Idiomatic Expressions
A Comparative Analysis of Compound Word Recognition Results between Human and Machine Annotation

Totok Suhardijanto, Zahroh Nuriah and Setiawati Darmojuwono
Department of Linguistic, Faculty of Humanities, Universitas Indonesia, Depok, Indonesia
totok.suhardijanto@ui.ac.id

Keywords: corpus, compound, compositionality, machine annotation.

Abstract: This paper presents our research progress in building an automatic recognition system for compound words in Bahasa Indonesia. Our goal is to develop a system that is able to distinguish significant multiword expressions and other insignificant groups of words. For instance, *rumah tangga* 'household' should be considered as a significant cluster of words rather than *rumah kayu* 'wooden house'. It is not easy to differentiate a compound word and an ordinary phrase in Bahasa Indonesia because there are no specific phonological markers like accent in German or Dutch. The orthographical markers are not always present, *rumah tangga* is written with a space while *kacamata* 'glasses' not. In this paper, we compare and analyze the results of machine and human annotation. The automatic annotation system is built with a statistical machine learning algorithm called conditional random field. Data for annotation task is collected from newspaper and magazine articles. In this analysis, the mixed method was applied to reveal the differences between human and machine annotation. The result showed that the machine still performed 69% of accuracy and had several error patterns in compound word recognition tasks. Human annotation is trivial due to personal annotator backgrounds.

1 INTRODUCTION

Electronic standard corpora of a language are very beneficial not only for linguistics but also for other social researches. There are several number kinds of corpora such as standard corpora, comparable corpora, or parallel corpora. The most important of those corpora to be used maximally is that the corpora are grammatically and semantically annotated so that the corpora can be easy automatically processed. For researchers, annotated corpora would ease them to dig up knowledge behind the texts.

Up to now, Indonesia has not got any sufficient electronic annotated corpus. There are several Part of Speech (POS) tagging systems developed by INACL (Indonesia Association for Computational Linguistics), but the application can not recognize compound yet. In this study we tried to develop an application to recognize idiomatic expression. With this recognition a language can be explored more precisely.

This paper presents our attempt to develop a part of speech tagging system that has an ability to recognize a compound word in Indonesian. The system is built based on the investigation of Indonesian compound words by using a corpus study. This research was conducted to determine the shortcoming in our automatic annotation system by comparing the results of human and machine annotation.

2 LITERATURE REVIEW

Until today, Indonesian linguistic experts have not reached an agreement on the definition of the Indonesian compound word. To determine a compound word, phonological criteria are often used which is not applicable in Bahasa Indonesia (Muhadjir, 1980: p.61; Badudu, 1980: p.16). In Bahasa Indonesia there are no specific phonological markers like accent in German or Dutch. The orthographical markers are not always present, *rumah tangga* is written with a space while *kacamata* 'glasses' not. In this section, morphosyntactic and semantics perspective of compound words,

Suhardijanto, T., Nuriah, Z. and Darmojuwono, S.
Teaching Machines to Recognize Idiomatic Expressions - A Comparative Analysis of Compound Word Recognition Results between Human and Machine Annotation.
DOI: 10.5220/0007167603760380
In *Proceedings of the Tenth Conference on Applied Linguistics and the Second English Language Teaching and Technology Conference in collaboration with the First International Conference on Language, Literature, Culture, and Education (CONAPLIN and ICOLLITE 2017) - Literacy, Culture, and Technology in Language Pedagogy and Use*, pages 376-380
ISBN: 978-989-758-332-2

particularly in relation to the development of compound word applications for Indonesian corpora, will be discussed. As Chaer (1980: p.48) noted during the symposium Tata Bahasa, which is held by the Lembaga Linguistik Fakultas Sastra Universitas Indonesia in celebrating Sumpah Pemuda 28 October 1979, that it is necessary to reanalyze Indonesian compound words based on big corpora.

More than twenty years ago, Kridalaksana (1980: p.32) argued that all elements of compound words must be a basic morpheme and at least one of them must be a bound morpheme. This opinion is still relevant in distinguishing compound words from word groups. Kridalaksana also stated that constructions that contain forms such as eka-, panca-, multi-, tuna- etc. are compound words, since those forms have no grammatical function, but they have lexical functions. Hence, it can be concluded that elements of the compound words must be a basic morpheme and at least one of them must be a bound morpheme which is not an affix. This definition cannot retain if we accept the concept potential words of Booij (1998: pp.52-53) since those morphemes have lexical functions.

Some Indonesian linguist (Badudu, 1980; Keraf, 1980; Kridalaksana, 1980) distinguished compound words morphosyntactically, namely that a compound word undergoes a derivation process such as reduplication and affixation as one unit. *Sapu tangan* 'handkerchief' and *tanggung jawab* 'responsibility' are compound words since the reduplication of *sapu tangan* is *saputangan-saputangan*, and the deverbal noun of *tanggung jawab* is *pertanggungjawaban*. But this is not always the case. Keraf (1980: p.59) noted that the form *sapu-sapu tangan* and *pertanggungan jawab* is also acceptable. Badudu (1980; p.15) mentioned another characteristic of compound words namely that they cannot be inserted by other element. A word like *rumah sakit* 'hospital' shall not be acceptable if it is inserted by *-nya* 'his', **rumahnya sakit*. Nevertheless, Kridalaksana (1980: p.26) argued that these words can actually be inserted by a preposition, for example *rumah untuk orang sakit*.

With regards to the identification of compound words from semantic perspective, there is still ambiguity on the concept of compound words' definition. For instance, the construction of compound word has a high degree of closeness, what is meant by "high degree of closeness" is vague, but the criterion "compound words do not refer to the referent of each constituent element but to a new referent" in terms of semantics can be interpreted that the combination of the words' meanings in a compound word has a united idiomatic meaning referring to a denotative meaning.

3 METHOD

In this research, the development POS annotation system with compound word recognition was conducted based on a conditional random field (CRF) algorithm that is derived from Hidden Markov Model (HMM). This algorithm is a class of statistical model method that applies structured prediction. It means that CRF algorithm uses discriminative approach. According to Sutton and McCallum (2011), CRF is a type of discriminative indirect probabilistic graphic model. In the implementation of CRF in POS tagging, it is necessary to determine first a set of *f*i feature function. In this algorithm, a POS label decision can be determined through its context, that is the preceding and following words.

In the comparative analyze, several corpus techniques were implemented to explore data. These techniques consist of word frequency list, collocation, and concordance (see Lindquist, 2011 and Cheng, 2011). The frequency list was applied to reveal common error patterns either in human or in machine annotation result. Furthermore, the collocation method was implemented to elaborate the error patterns. Finally, it was also applied to find out the contexts of erroneous annotation in the corpus.

4 FINDINGS AND DISCUSSION

In dealing with compound word recognition from morphosyntactically perspective, a semantic characteristic can be very useful. A compound word refers to one referent that acts as unit that cannot be inserted by another element. The argument of Kridalaksana with the example *rumah untuk orang sakit* cannot retain since *rumah putih* 'white house' as a phrase can also not be inserted by the preposition *untuk* 'for' (**rumah untuk putih*). It is not only a matter of insertion, but it is a matter of what you insert. Consider the following examples:

rumah sakit	*rumah putih*
**rumahnya sakit*	*rumahnya putih*
perumasakitan	**perumahputihan*

Based on the characteristics of compound words as stated by Suwarso (1980), compound words and word groups can be distinguished. According to Suwarso (1980: p.38) groups of words can be identified by the nature of the word relationships: the

attributive relationship and the coordinative relationship (equal relationship). Attributive relationships are specified as allocative (*rumah makan*, 'restaurant'), instrumental (*meja tulis,* 'writing desk'), possessive (*hulubalang* 'guard'), final/aim (*bina marga* 'road development'), partitive (*luar negari* 'overseas'), ablative (*orang Jawa* 'Javanese people'), comparative (*merah jambu* 'pink'), quantitative (*penuh sesak* 'crowded').

After the distinction between word groups and compound words have been established, the next challenge is to find criteria to differentiate idiomatic relations among compound words. Idioms can take various forms, from word to sentence, e.g. *tikus kantor* 'office mouse (=corruptor)', *suara emas* 'golden voice (=good singer)', *besar kepala* 'big head (=vain)', *badan Amran setelah sakit tinggal tulang berbalut kulit* 'After being sick, the body of Amran is only bone and skin (=Amran is now getting slim)'.

Compound words can be distinguished from semantic perspective. For instance, an idiomatic phrase *kambing hitam* is a full idiom, because the meaning of the idiomatic phrase cannot be traced from the meaning of its element and the new meaning/change of meaning is unrelated to the meanings of the elements (opaque). *Kambing hitam* can mean (1) black goat (a group of words) or (2) a person who is blamed (idiom). Therefore, *kambing hitam* could be a phrase or compound words with an idiomatic meaning. Semi idioms can be recognized by the meaning of one of its constituent elements, such as *daerah hitam* 'black area', the meaning of *daerah* still refers to a place but the meaning of *hitam* changes from a colour to an environment where people commit a crime, prostitution etc. Based on these examples it can be concluded that, idiomaticity can be regarded as a feature of compound word in Bahasa Indonesia.

In relation to the need for tagging to identify word groups and compound words, the following steps are proposed:

- consists of more than one word;
- Is there attributive or coordinative relationship? If yes, it is a group of words, if not there is a possibility of compound words;
- (3) Idiomatic construction shows a high degree of closeness so that it is an integral part and its elements cannot be replaced with another element (e.g. *duta besar* cannot be substituted by *duta kecil**, or *duta tua** etc, *duta besar* means ambassador). One element or both of them have metaphorical meaning (e.g. *besar* is not a size but its meaning is idiomatic).
- If points 2 and 3 are the case, the word construction is a compound word.

In this research, once the development of the POS tagger was accomplished, we conducted an experiment to apply the tagger onto our training corpus. The experiment was applied to a corpus with 2 million words.

Along with the machine annotation experiment, we also conducted a human annotation experiment on the same corpus. The results of the POS and compound annotation then were comparatively analyzed by using corpus methods. The use of corpus methods is chosen to make the comparative analysis more practical, robust, and fast.

In this research, we conducted an experiment with dataset consisting of 2 million corpus. The result of annotation system showed that the machine annotation accuracy reached 69,229%. The reference set is a work of compound word annotation done by experts in Indonesian linguistics. Among the correct annotated results, some recognized compound words are related to multiple expression words (MWEs) with high frequencies in Sketch Engine (https://www.sketchengine.co.uk/). It means that the more frequent a compound word in a corpus the more identifiable by machine annotator is.

Aside from the experiment with machine annotation, in this research we also set a group of students to do an annotation task with the same corpus. As predicted before, the result of human annotation is much higher than those of the machine annotation. However, in few cases, when human annotators made a mistake by mislabelling a phrase as a compound word, e.g. *jalan raya* 'main road' (*jalan* 'road'; *raya* 'big, large, main'), the machine left it unlabelled.

In the experiment, there are 92 erroneous annotated results that can be classified into four different types of errors: incompleteness, miscategorization, contextual error, and other. Incompleteness refers to annotation errors due to non-completion of annotation, e.g. (*bekerja*, VB) 'to work', (*sama*, COMP) 'together', > *bekerja sama* (COMP) 'to collaborate'. In the example, the machine only labelled *sama* correctly, but the other component *bekerja* was failed to be recognized.

Miscategorization is an error due to a failure in recognizing a compound word. This error consists of two different types that is over identification and under identification. Over identification refers to a situation when the machine recognized a compound word to excessive degree, for example: ('*jangka*', 'COMP'), ('*pendek*', 'COMP'), > not labelled. In this case, the machine should not categorize *jangka pendek* as a compound word, but rather as a phrase because of its compositional meaning. Meanwhile, under identification refers to a contrast situation to

over identification when the machine was unable to identify a compound word, e.g. ('*minuman*', 'NN'), ('*keras*', 'JJ'), > *minuman keras* (COMP). In the example, all components *minuman* and *keras* should be tagged as COMP (compound word), not as NN (noun) and JJ (adjective) respectively.

Contextual errors refer to an annotation mistake due to a surrounding environment, for example, a structural pattern which occurred in a passage or sentence. There are four types of contextual errors found in data: pattern-related, serial verb, proper noun, and sequential error.

The first type is the pattern related error that shows a correlation with a grammatical pattern. In this case, it is a collocation or compound word pattern, e.g. ('*meninggal*', 'COMP'), ('*mendadak*', 'COMP') > should not be tagged as a compound. This error occurred several times in data so it means that this error occurred in relation to the word *meninggal* that is closely related to a compound word pattern of *meninggal dunia*.

The second type is similar to the first type, but the pattern is related to a serial verb. For instance, ('gagal', 'VB'), ('bayar', 'COMP') > should not be labelled as a compound because *gagal bayar* 'fail to pay' is a serial verb.

The third type is an annotation error that related to a proper noun. This error occurred when the machine wrongly recognized a proper noun as a compound word, for example, ('*Baldwin*', 'COMP'), ('*Lonsdale*', 'Z') > should be annotated as a proper noun or NNP.

The fourth type is an annotation that is wrongly implemented to adjacent words around or next to a compound verb. For example, ('*gembira*', 'COMP'), ('*kerja*', 'COMP'), ('*sama*', 'JJ'), > annotation should be applied to *kerja sama* 'cooperation', and not to *gembira kerja* that means nothing.

Table 1: Types of annotation errors.

Type	Frequency	Percentage
Incomplete	80	17,97752809
Miscategorization	230	51,68539326
Contextual Error	130	29,21348315
Other	5	1,123595506
TOTAL	445	100

Table 2: Subtypes of miscategarization errors.

Type	Frequency	Percentage
Over identification	165	71,73913043
Under identification	65	28,26086957
TOTAL	230	100

Table 3: Subtypes of contextual errors.

Type	Frequency	Percentage
Pattern Related	50	38,46153846
Serial Verbs	20	15,38461538
Sequence Related	45	34,61538462
Proper Nouns	15	11,53846154
TOTAL	130	100

Table 1, 2, and 3 show the number of errors that are related to annotation. Table 1 shows that the most significant error is the miscategorization error with 51,685% of the total error number. Among miscategorization errors, over identification errors are the most frequent phenomena. Over identification errors are characterized by a wrong annotation of a high frequent phrase, so it indicates that over identification is related to a statistical significant collocation. It means that the machine ability to distinguish between a compound word and a statistical significant phrase such *jangka pendek*.

5 CONCLUSIONS

The big problem with regard to compound words in Bahasa Indonesia is that there are no adequate grammatical and phonological and orthographical markers with regard to compound words. For this reason, it is quite tough to build an automatic annotation for a compound word in Indonesian.

In this research, the annotation system was built based on a conditional random field algorithm. This system was implemented to a 2 million dataset to evaluate the results. The machine performed quite well with medium accuracy (more than 69 %) in recognizing compound words.

Among annotation errors, the most dominant error is miscategorization. It means that the machine has a problem in classifying and clustering a group of words whether it is belong to a compound word or just a phrase. In the future, the algorithm needs to be reviewed to improve its performance.

REFERENCES

Badudu, J. S. 1980. Kata Majemuk dalam Bahasa Indonesia, in Masinambouw, E.K.M., *Kata Majemuk: Beberapa Sumbangan Pikiran*. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 13-16.

Booij, G. A. S. 1998. *Morfologie: De woordstuctuur van het Nederlands*. Amsterdam: Amsterdam University Press.

Chaer, A. 1980. Usaha Mencari Identitas Kata Majemuk dalam Bahasa Indonesia, in Masinambouw, E.K.M.,

Kata Majemuk: Beberapa Sumbangan Pikiran. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 13-16.

Chen, W. 2011. *Exploring Corpus Linguistics: Language in Action*. London: Routledge.

Keraf. G. 1980. Kata Majemuk, in Masinambouw, E.K.M., *Kata Majemuk: Beberapa Sumbangan Pikiran*. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 53-60.

Kridalaksana, H. 1980. "a+b = ab", in Masinambouw, E.K.M., *Kata Majemuk: Beberapa Sumbangan Pikiran*. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 25-36.

Lindquist, H. 2009. *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Muhadjir. 1980. Beberapa Ciri Kata Majemuk, in Masinambouw, E.K.M., *Kata Majemuk: Beberapa Sumbangan Pikiran*. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 61-66.

Sutton, C., McCallum, A. 2011. An Introduction to Conditional Random Fields for Relational Learning. Accessed in September 30, 2017, in https://arxiv.org/pdf/1011.4088v1.pdf

Suwarso, S. 1980. Kata Majemuk dalam Bahasa Indonesia, in Masinambouw, E.K.M., *Kata Majemuk: Beberapa Sumbangan Pikiran*. Jakarta, Fakultas Sastra Universitas Indonesia, pp. 37-40.