Online version available in : http://arbitrer.fib.unand.ac.id

JURNAL ARBITRER

| 2339-1162 (Print) | 2550-1011 (Online) |

Article

Computational Linguistics Models and Language Technologies for Indonesian

Dian Noviani Syafar[1], *Ria Febrina*[2]

[1] *English Education Department, STKIP PGRI Sumatera Barat, Indonesia*
[2] *Indonesian Department, Universitas Andalas, Indonesia*

SUBMISSION TRACK

Recieved: March 28, 2017
Final Revision: May 03, 2017
Available Online: May 15, 2017

KEYWORD

Computational linguistics, Indonesian, spelling, translation.

KORESPONDENSI

E-mail: author@email.com

A B S T R A C T

The purpose of this research is to describe computational linguistics as a study of science that should pay full attention to linguistics researches improvement. The type of research is literature review and experimental research by designing a software model for Bahasa. The result of the research shows that computational linguistics is a field of linguistics that can be used as a solution to overcome a problem related with the spelling correction and grammar for language users. This field of linguistics is related to software engineering designed to educate the public in producing languages, it can be Bahasa, regional language, and English language as the foreign language for Indonesian. The public can know the standardization of writing a language and equivalence translation between the target language and the source language that can be also precisely acquired. In addition, the writer provides several practical examples on how computational linguistics can be applied to the development of writing skills. For instance, the concordance enables to see any word or phrase in context so that one can see what sort of company it keeps. Thus, the users can, for example, see the correct form based on Bahasa Indonesia rules between the words which they often confuse (e.g., *gadget* vs. *gawai*).

I. INTRODUCTION

With the advances of globalization and technology, it has been estimated that over a billion people are learning several languages in various parts of the world and the numbers are growing, not only for English but for other languages as well. Spoken by over a quarter of the world's population and enhanced by the presence of internet, it has become the operating system of the global communication. Indeed, there is estimated 375 million English as a second langguage (ESL) and 750 million English as a Foreign language (EFL) learners around the world (Graddol, 2006).

Moreover, there are many researchers have done projects on translation strategies as a tool for teaching future translation process by adapting computer assisted translation tools. Some of them also develop computerized

DOI: https://doi.org/10.25077/ar.6.1.45-52.2019

databases that provide translators with textool reference and consultation material in order to maintain coherence and consistency in their decision making during the translation process (Candal-Mora & Vargas-Sierra, 2013).

Klimova (2013) reveals how corpus linguistics can be applied in the development of writing skills by carrying out specialized software programmes on a computer. WordSmith Tools is one of program mentioned for checking at how words behave in English texts. For foreign language teachers and learners, this program gives many advantages to see any word or phrase in context since they see the diffrences between difficult words (e.g. excited vs exciting) then disecover the most appropriate to use.

Meanwhile, related problem in Bahasa, the correction toward social media texts and data collection corpus that come from conversation subtitles in the movies are also recently presented (Kusumawardani, R. P, et.al, 2018, Chowanda, A, 2018) Chowanda A also collects 1961 movies with Indonesia subtitles were collected dan precessed, leaving 1678320 unique words from more than 24M words in Bahasa in the dataset. Then, another rsearch creates a neural word embeddings using word2vec trained on over million social media messages representing a mix of domains and degrees of linguistic deviations from standard Bahasa.

For Indonesian language or *Bahasa Indonesia*, English seems to give influence Indonesian with a great deal, especially with its loanwords due to Indonesians tends to associate the paradigm of high status and prestige. With the prominent role of Indonesian, the Indonesian government has made various efforts to maintain the language, especially through its National Center for Language Planning (*Badan Pengembangan dan Pembinaan Bahasa*). It also affects the development of Indonesian, so that it makes "increasingly less room" for the use of National Language (Badan Bahasa Kementrian Pendidikan dan Kebudayaan, 2015). Such a sociolinguistic factor, English is considered to be a symbol of social pride. This factor seems to encourage the unneeded use of English in Indonesian advertisements and media despite their frequent inaccuracies, for example, a well-known fast-food franchise in Indonesia wrote *safety your hunger* to mean *satisfy your hunger*. Although English is only a foreign language in the country, this increasing use of the language is often seen as a threat to the maintenance of Indonesian as the country's national language and symbol. The government has made rules to control the pervasive effect of English loanwords on the national language, but they do not seem to work well because Indonesian people may not be aware of or simply ignore the rules.

These language users provide a burgeoning market for tools that help identify and correct learners' writing errors. Unfortunately, the errors targeted by typical commercial proofreading tools do not include those aspects of a second language that are hardest to be learnt. Evidently, automatic grammar checkers are much needed to help learners improve their writing. However, typical English proofreading tools do not target specifically the most common errors made by second language learners.

The concern here is apparently beyond how much English terms are used by Indonesian users and communities from various aspects. In addition, the use of English is not only restrictively applied in the realm of education, but also broadly spread to become a major factor in the development of language industry nowadays. Every industry player consciously uses language to promote each activity that carried out, either on banners, billboards, pamphlet, or other commercial service advertisements. The promotion is done to influence the community to choose their own products and ignore other similar products of equal quality.

This paper gives an overview of scientific program Computational Linguistic models and language technologies for Indonesian. It

also presents the role of computational linguistics in constructing the English forms that find difficult language users find most difficult -- constructions containing correct spelling of words or phrases. It provides an overview of the automated approaches that have been developed to identify and correct these learner errors in a number of languages.

Furthermore, related to the importance of foreign languages in publishing scientific works, Indonesian also have experienced difficult things. The ability of Indonesian to acquire foreign languages is very limited. Meanwhile, the need for translation is very high. Therefore, a number of Indonesian rely on translation machines to translate Bahasa scientific works into English scientific works. However, the existing of translator machine apparently has limited translation so the results that achieved through the machine do not match with the source language. One of the cause of discrepancies is the lack of involvement of linguists in determining translation models that are suitable for both languages, source language and target language.

Based on this case, an appropriate strategy is needed to overcome the problems in the language industry in Indonesia related to Bahasa, regional languages, and foreign languages. One strategy that can be used is using technology in which the technology itself influences community in carrying out an action quickly. Second, people can receive information simultaneously from one sender. This broad dissemination of information can be used to overcome the problems for language users. Indonesian need systematic language education to overcome the errors in producing the language itself.

II. METHODS

The problems solving in the language will be done by conducting literature studies related to language research conducted by a number of experts. The language research focused on language research related to computational linguistics. A number of language studies in the field of computational linguistics will be described, then it will be analyzed to find the right formula in producing solutions for the language users.

Since computational lingusitics focuses n the study of language uses aided by computers, there exists a link to library and information science. In this research, the literature studies will be assisted with a simple research on computational linguistics, specifically related to the presence of applications that can help the community in overcoming the writing errors in Bahasa. The application is generated through a software engineering process that has gone through analysis tests and has been applied in a simple prototype.

The methods used in this study are divided into three stages, namely the method of collecting data, the stage of data analysis, and the stage of presenting the results of the analysis. At the stage of data collection, the data are collected by applying library techniques, using written sources that taken from computational linguistic journals. As stated by Zaim (2014: 95), there are three articles relating with computationall inguistics are conducted in this study which deal with English translation, Javanese translation, and Indonesian spelling correction.

Next, domain analysis is used by Spradely (1980) at the stage of data analysis. At this stage, the data collection are analyzed by sorting data related with the use of computational linguistics in language research, then connect and explain the importance of computational linguistics for Indonesian users, especially in using mother tongue, national language, and foreign languages according to the correct rules . At the stage of presenting the results of data analysis, informal presentation methods are used, the data are explained descriptively by using sentences.

III. RESULT

As a manifestation of community involvement in producing language as an economic product, it can be seen in an economic vehicle belonging to the microeconomic community. The various kind of communities need for fast food marketing causes the micro-economic community mobilize their business in this field. To fulfill market's needs, microeconomic communities produce the food products and produce languages to attract the buyers. Because of the low level of education, the community relies on the sound of language to produce products that come from foreign languages.

Consider, *sandwich* is pronounced *[senwic*] were finally produced as *sunwich* on their operational vehicles. As well, the *nuggets* that were heard [nagget] were produced as *negget* on their operational vehicles. At this stage, the community is trying to become an international community by producing English words.

In other micro-businesses sector, a number of communities also rely on the sound of language to produce Bahasa products. Public service advertisements that invite people to consume local fruits then move the micro business community to sell local fruits. Because of the community's needs for fruits are getting higher, the micro business community provides all kinds of fruits from their own agricultural products and from overseas agricultural products. To promote and increase the number of buyers, the micro business community compensates for economic products by providing advertisements or billboards. With the local fruit campaign, the micro business community then produced billboards with the writing that they provided *buah lokal* and *buah interlokal*.

At this stage, the language community seeks to present word pairs for their economic products. However, the limitedness understanding of language causes that they produce the wrong language. The pair of words for *buah lokal* that should be *buah impor*, it is actually written with *buah interlokal*. Related to the language production, it is not only weak economic people who produce it, middle-class society and upper class society also make mistakes in producing language. One of them is language production on a doctor's billboard. Every doctor does not check the language produced on the billboard so that a nameplate is found in the form of a *praktik dokter* and a nameplate in the form of a *praktek dokter*. From the two signboards, only one signboard is in accordance with the rules of writing in Bahasa, namely *praktik dokter*.

Not only that, for the writing of each doctor's degree, they also varied in producing language, namely *dr., Dr.,* and *DR*. In Bahasa, the writing of the three has different functions. The only billboard that became the correct language production was written in *dr*. Language production in the industrial world carried out by various layers of community certainly it can affect the use of language by the community itself. The errors in producing language will also affect the community in using wrong language in an ongoing. In fact, in Article 36 of the 1945 the Constitution, it has been stated that "State language is Bahasa" (Ahmad, 2007). Thus, using the correct Bahasa is a must for Indonesian.

The use of the correct Bahasa has been standardized by Badan Bahasa Republik Indonesia through two things, namely Law No. 50 of 2015 concerning Pedoman Umum Ejaan Bahasa Indonesia (PUEBI) and Kamus Besar Bahasa Indonesia (KBBI). However, not many people use the standardization in producing Bahasa. In fact, the presence of editors and linguists in various companies move in the language industry was also eliminated. This is a major cause in the emergence of many language errors in various public facilities nowadays.

In the next stage, related to the importance of foreign languages in publishing scientific works, Indonesian also experienced difficult things. The ability of Indonesian to acquire

foreign languages is very limited. Meanwhile, the need for translation is very high. Therefore, a number of Indonesian rely on translation machines to translate Bahasa scientific works into English scientific works. However, the existing of translator machine apparently has limited translation so the results that achieved through the machine do not match with the source language. One of the cause of discrepancies is the lack of involvement of linguists in determining translation models that are suitable for both languages, source language and target language.

Based on this case, an appropriate strategy is needed to overcome the problems in the language industry in Indonesia related to Bahasa, regional languages, and foreign languages. One strategy that can be used is using technology. Technology influences community in carrying out an action quickly. In seconds, people can receive information simultaneously from one sender. This broad dissemination of information can be used to overcome the problems for language users. Indonesian need systematic language education to overcome the errors in producing the language itself.

Computational linguistics is concerned with a system of words or symbols that can be communicated to a computer, especially to enter computer instructions through words that are easily understood and then translated into machine code. Computational linguistics can be a field of science used to help solve the problems in the language industry nowadays. The use of information systems in the form of software engineering in helping the field of linguistics is the only way to reach people who are now in contact with cyberspace. Technology access that can reach all levels of community quickly, especially people who are literate.

Integration of Authoring tools and Translation Memory System

The application of computational linguistics to overcome the problems in the language industry is done through software engineering. According to Sommerville (2011, 7–8), software engineering is a scientific discipline that discusses all aspects of software production, from the beginning stage of system specifications to system maintenance after it is used. Software engineering must be designed to work to overcome problems for language users.

Software engineering in the field of linguistics is not only related to the technical processes of software development, but also related to the activities, such as software project management and the development of linguistic theory to support the production of software. Related to software management, software evaluation and improvement will be carried out to help to create effective and efficient software.

Systematic and organized software engineers can produce high-quality software. Related to that case, Sommerville (2011, 36--41) stated that it is needed to do some software processes. The software process is a series of activities and the relevant results that produce software. There are four basic process activities that are common to all software processes, namely (1) specifying software and limiting its operation, (2) designing and implementing software, (3) validating software so that it can be known that the software works in accordance with customer's needs, and (4) evolving software so that it can develop to deal with the changing of customer's needs.

Rösener (2010: 1) stated that computational linguistics can be used to solve language translation problems. In his research, Rösener (2010) integrates a tool into a *database*-based translation system. This system allows the translators to maintain the terminology of the two languages because in the tool stored the specific information about vocabulary, gender, and structure.

The system developed by Rösener (2010), it is not only allow the users to translate the

languages, such as grammar and language styles in the source language and target language, but it also can help in writing according to the source language and target language. This translation is used to maintain the quality of the text so that the translation language matches with the language of the original document.

Transliterator : Jawatex

Utami et al. (2013: 78) built a transliteration model named JawaTeX to translate Latin text documents into Javanese characters. From the results of the tests, Utami et al. (2013) prove that the users can write every word or term correctly, including absorption words according to the original pronunciation. In fact, users can write or rearrange the Latin spelling in the source text

Utami et al. (2013: 92) described a number of capabilities of JawaTex in translating Latin text into Javanese characters, namely (1) can translate more than one sentence, even paragraphs; (2) can translate Latin *string* based on the pattern of decapitation of Latin *string* that search without transcription; (3) can search for word similarity to correct word spelling errors using two English and Bahasa vocabulary lists; (4) can do writing formatting; (5) can handle the use of Latin characters that do not have alphabetical equivalents in Javanese script; (6) can handle the use of diphthongs; (7) can handle the use of Roman numbers; (8) can handle the ambiguity of the use of dots as word end markers, abbreviated markers, and decimal markers; (9) can handle the ambiguity of the use of spaces, quotation marks (`), and hyphens (-); (10) can handle the use of more than three consonant characters; (11) can encode the mapping pattern and produce a transliteration pattern in accordance with the list of model mapping patterns that have been compiled; (12) can carry out transactions with the placement of Javanese script in accordance with the writing scheme that has been designed; (13) can make various changes in the form of Javanese characters depending on their position in a sentence, as well as other characters who follow and follow it; (14) can be implemented on several operating systems, and has been tested on Windows and Linux; and (15) can be implemented through the website, and tested at http://www.jawatex.org. Looking at the JavaTeX design to help translate the Latin text into Javanese script, the concept that is built can be used as a basis to be developed on any type of language script.

Spelling Correction in Indonesian : ejaan.id

Febrina, Hilman, and Abilowo (2018) also use computational linguistics to produce Bahasa writing models to produce the Bahasa rules. They created the *ejaan.id* page to help Indonesian improve their scientific work in Bahasa. The results of the research conducted showed that every Bahasa scientific work contains non-standard words that are not in accordance with the KBBI. Therefore, the *ejaan.id* page is designed automatically to correct the non-standard words into standard words.

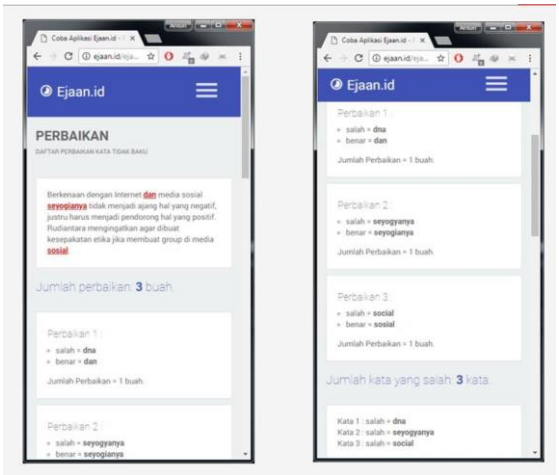

Fig 1: Ejaan.id

Febrina, Hilman, and Abilowo (2018) explained that when a user put a number of words, sentences, paragraphs, or discourses into the box provided on the *ejaan.id* page, non-standard words immediately turn red. Under the box, a box is provided which has presented the results of the improvement from the non-standard word, as well as incorrect data and correct data are presented.

Through the improvement of non-standard words through technology-based applications using the *ejaan.id* page, Bahasa users will be notified that the word used in the paper is not in accordance with the rules for using Bahasa. Furthermore, through the page, the users are also notified of words that are in accordance with the rules of use of the Bahasa. With this application, Indonesian users are expected to be able to limit the use of non-standard words or words that are not in accordance with Bahasa standards, both in written and in oral variety. Furthermore, the users can take the results of these improvements as scientific papers that have been edited by technology-based editors. Editing is done by using standardization in the form of Law No. 50 of 2015, namely PUEBI and also adapted to the KKBI.

Through the development is done through computational linguistics, it is expected that Bahasa users amount to 183 million working age population (more than15 years) can use it to produce the correct language in accordance with the rules and grammar.

According to the Badan Pusat Statistik (2016), angka melek huruf (AMH) of the Indonesian population aged more than 15 years is 92.37 percent. So, every 100 people aged more than 15 years, 92 are literate. The literate population certainly needs to know the standardization of writing a language and translating from the source language to the target language.

IV. CONCLUSION

Based on the research that has been done with computational linguistics, it can be concluded that the field of computational linguistics is a field of linguistics that can be used as a solution to overcome the problems for language users. The field of linguistics and the field of information system science can be collaborated to produce engineering products in educating the community related to the production of languages among Bahasa, regional languages, and foreign languages. Grammatical and spelling error detection for language users has been an area of active research which involves pinpointing some words in a given sentence as grammatically erroneous and possibly offering correction. Therefore, language users or communities the standardization of writing a language and it can get the translation of the target language according to the source language. To conclude, a new method for correcting serial errors in a given words or phrase in learner's writing has been introduced.

REFERENCES

A Ahmad, Sabri. 2007. "Undang-Undang Dasar 1945 Pasal 36". Jakarta: Quantum Teaching, Ciputat.

Badan Bahasa. 2017. *Ejaan Bahasa Indonesia*. Jakarta: Kementerian Pendidikan dan Kebudayaan Indonesia.

Badan Pusat Statistik. 2016. "Angka Melek Huruf", https://sirusa.bps.go.id/index.php?r=indikator/view&id=313, retrieved on September 11, 2018 at 02.36 pm.

Bolshakov, Igor, Alexander Gelbukh. 2004. Computational Linguistics : Models, Resources, Application. Mexico: Instituto Politecnico Nacional.

Febrina, Ria, et.al. 2018. "*Ejaan.id* sebagai Inovasi Pengeditan Naskah Berbasis Komputerisasi". Makalah untuk Prosiding Kongres Bahasa XI Badan Bahasa Republik Indonesia.

Gamon, Michelle. Using mostly native data to correct errors in learners' writing. In *Proceedings of the Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Los Angeles, 2010.

Graddol, David, 2006. English next: Why global English may mean the end of 'English as a Foreign Language.' UK: British Council.

Rachele De Felice and Stephen G. Pulman. Automatic Detection Of Preposition Errors In Learner Writing. *CALICO Journal*, 26(3):512–528, 2009

Rösener, Christoph. 2010. "Computational Linguistics in the Translator's Workflow". Proceedings of the NAACL HLT. Los Angeles, California.

Sommerville, Ian. 2011. *Software Engineering (Rekayasa Perangkat Lunak)*. Jakarta: Erlangga.

Utami, dkk. 2013. "Penerapan *Rule Based* dalam Membangun Transliterator Jawatex", Jurnal Berkala MIPA No. 23 Vol. 1 Edisi Januari.

Wibisono, Setyawan. 2013. "Aplikasi Pengolah Bahasa Alami untuk Query Basis data Akademik dengan Format Data Xml". Jurnal Teknologi Informasi DINAMIK Volume 18, No.1, Edisi Januari.