A Gold Standard Dependency Treebank for Indonesian

Ika Alfina, Arawinda Dinakaramani,
Mohamad Ivan Fanany, and Heru Suhartanto
Faculty of Computer Science, Universitas Indonesia
Depok, Indonesia
ika.alfina@cs.ui.ac.id, arawinda.dinakaramani@ui.ac.id,
{ivan, heru}@cs.ui.ac.id

Abstract

Resources for syntactic parsing for Indonesian are very limited, as there are only two dependency treebanks publicly available and both are small in size. Not only that, we found out that the word segmentation method used by both treebanks needs improvement. Therefore, in this work we proposed a revision for one of these treebanks, Indonesian Parallel Universal Dependencies treebank. Besides improving word segmentation, we also improved POS tagging and syntactic annotations. Because in Indonesian grammar there are some special structures, we also proposed how to adjust UDv2 annotation guidelines with those Indonesian grammar rules. To evaluate the quality of the new treebank, we built Indonesian dependency parser model using Parsito (UDPipe) parser. Using ten-fold cross-validation, the model that built using the revised treebank had UAS of 83.33% and LAS of 79.39%, over the original treebank with UAS of 73.32% and LAS of 65.98%.

1 Introduction

Indonesian is a language spoken by more than 260 million people in 2019, but its resources for Natural Language Processing (NLP) research are still limited. Especially for the syntactic parsing studies, the availability of syntactic corpora (treebank) is scarce.

As far as we know, there are only two dependency treebanks for Indonesian that available publicly. Both are provided by the Universal Dependencies (UD)[1]. The first one is UD Indonesian-GSD (McDonald et al., 2013) that consists of 5,593 sentences, and the second one is UD Indonesian-PUD (Zeman et al., 2018) that consists of 1,000 sentences.

Unfortunately, after conducting reviews to the quality of both treebanks, we found out major flaws, especially in the word segmentation that does not comply with Indonesian grammar. In UD Indonesian-GSD, all clitics are not separated from the main words. While in UD Indonesian-PUD, words with clitics are always be split in any context and the reduplicated or hyphenated words that occur frequently in Indonesian are always separated into multiple tokens.

In this work, we proposed a revision to the UD Indonesian-PUD since its size is smaller than the UD Indonesian-GSD. Meanwhile, we also observerd there are some characteristics of Indonesian grammar that needs special treatments. UD created guidelines for cross-linguistically grammatical annotation. The current version of the annotation guidelines is named Universal Dependencies v2 (UDv2). To address special constructions in Indonesian grammar, we proposed some adjustments to UDv2 guidelines.

The contributions of our work are two folds. First, we proposed some adjustments to UDv2 annotation guidelines to build dependency treebank for Indonesian. Specifically, we proposed special treatments in word segmentation and POS tagging process for Indonesian and proposed the use of some dependency relations for certain language constructions in Indonesian grammar. Second, we proposed a revision to the UD Indonesian-PUD treebank of 1,000 sentences, resulting in a better gold standard depen-

[1] https://universaldependencies.org/

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 1-9, Hakodate, Japan, September 13-15, 2019

dency treebank for Indonesian. This revised treebank had been made public[2].

The rest of this paper is organized as follows: Section 2 addresses the Indonesian grammar; Section 3 describes the Indonesian PUD treebank; Section 4 explains the proposed annotation guidelines; Section 5 describes the annotation procedure and the statistics of the revised treebank; Section 6 discusses the experiments and results, and finally, Section 7 presents the conclusions and future work.

2 Indonesian Grammar

In this section, we discuss some Indonesian grammar rules that are relevant to our work in revising the UD Indonesian-PUD treebank.

2.1 Reduplicated Words

Some words in Indonesian are formed using reduplication (Sneddon et al., 2010). For example, the plural nouns such as *anak-anak (children)*, singular nouns such as *arak-arakan (procession)*, verbs such as *merobek-robek (shredding)*, adjectives such as *hiruk-pikuk (noisy)*, and adverbs such as *terus-menerus (continuously)*.

This characteristic implies that in the word segmentation process, this kind of words should not be split into multiple tokens.

2.2 Indonesian Clitics

A clitic is "*a morpheme in morphology and syntax that has syntactic characteristics of a word, but depends phonologically on another word or phrase*"[3]. Indonesian has two kinds of clitic: 1) As a personal pronoun; and 2) as a particle (Alwi et al., 1998). The clitics of personal pronoun are *ku- (I), -ku (me/my), kau- (you), -mu (you/your)*, and *-nya (him/her/it)*, and of the particle are *-lah, -kah*, and *-tah*.

As a clitic has a syntactic role, in the word segmentation process, we need to separate them from the main word. Furthermore, Larasati (2012) reported that by handling the clitic, the accuracy of an Indonesian to English machine translation system was improved.

Most of Indonesian clitics of personal pronoun have an ambiguous nature, especially *-nya*. Table 1 shows examples of words that ended with *-nya*. On this table, we use the UDv2 part-of-speech (POS) label[4] to abbreviate the word class. For each word, the syntax of how the word is formed and the actual POS are presented.

Word	Syntax	Actual POS
bukunya (her/his book)	NOUN + *-nya*	NOUN + DET
bukunya (the book)	NOUN + *-nya*	NOUN + DET
akhirnya (finally)	NOUN + *-nya*	ADV
khususnya (especially)	ADJ + *-nya*	ADV
jauhnya (the distance)	ADJ + *-nya*	NOUN
cantiknya (very beautiful)	ADJ + *-nya*	ADJ + ADV
sebenarnya (actually)	se + ADJ + *-nya*	ADV
dibukanya (open by him/her)	VERB + *-nya*	VERB + PRON
dibukanya (the opening)	VERB + *-nya*	NOUN

Table 1: Examples of the ambiguity of *-nya*.

We can see that the syntax of "NOUN + *-nya*" has three possible interpretations: "*-nya*" as the possessive determiner; "*-nya*" as the determiner; and "*-nya*" changes a NOUN into an ADV. While the syntax of "ADJ + *-nya*" has three possible meaning: "*-nya*" could change an ADJ into an ADV or a NOUN, or the meaning of *-nya* becomes "very" (ADV). The syntax of "*se* + ADJ + *-nya*" forms an ADV. The last syntax, "VERB + *-nya*" has two possible interpretations: "*-nya*" as the PRON following the VERB, or "*-nya*" changes a VERB into a NOUN. The use of "*-nya*" to change an ADJ or VERB into a NOUN is called the predicate nominalization (Sneddon et al., 2010, p311).

This shows how challenging it is to decide whether a token that ended with "*-nya*" should be split or not in the word segmentation process. It requires the information of POS tags of some words around the token with "*-nya*".

2.3 Compound Words

A compound is *"a combination of two simple words which come together to form a complex word"* (Sneddon et al., 2010). In Indonesian there are three ways to write the compound words: 1) as a single token, such as *kacamata (eyeglasses), matahari (sun)*; 2) hyphenated, such as *pemuda-pemudi*

[2]https://github.com/ialfina/revised-id-pud/

[3]https://en.wikipedia.org/wiki/Clitic

[4]https://universaldependencies.org/u/pos/index.html

(youngsters); and 3) as two tokens, such as *sapu tangan (handkerchief)*. Beside of noun compound, there are also compound words of verb, adjective, and so on. Table 2 shows some examples of Indonesian compound words.

POS	Examples
NOUN	*tanggung jawab (responsibility)*
VERB	*bertanggung jawab (to be responsible)*
ADJ	*luar biasa (excellent)*
ADV	*sering kali (often), kadang kala (sometime)*
NUM	*salah satu (one, a/an)*
DET	*salah seorang (a/an, for person)*
SCONJ	*di mana (where)*

Table 2: Examples of compound words in Indonesian.

Since a compound word has a syntactic role, we suggest that the compound words that have already written as a single token or hyphenated do not need to be split in the word segmentation process. While compound words written as two words need special treatment so that the relation between those two words is retained.

2.4 Noun Phrases in Indonesian

A noun phrase is *"a sequence of words which functions in the same way as a noun"* (Sneddon et al., 2010). A noun phrase always contains a noun as its head. There are two kinds of dependency direction, either head-initial or head-final (Hawkins, 1990). While English usually uses head-final direction for noun phrase, Indonesian mostly uses head-initial with some exceptions (Alwi et al., 1998). Table 3 shows some syntactic constructions of Indonesian noun phrases and their respective head directionality.

Type	Direction
NOUN + Demonstrative DET	head-initial
Quantity DET + NOUN	head-final
NOUN + Possessive DET	head-initial
NOUN + ADJ	head-initial
NOUN/PROPN + NOUN/PROPN	head-initial

Table 3: Head directionality of several types of noun phrase in Indonesian.

The following are some examples of Indonesian noun phrases:
1) *buku ini (this book)*
2) *dua mahasiswa (two students)*
3) *beberapa masalah (some problems)*
4) *rumah baru (new house)*
5) *rumah sakit (hospital)*
6) *rumahku (my house)*
7) *ekor anjing (the dog's tail/tail of the dog)*
8) *rumah Ika (Ika's house)*
9) *pemilik toko (a store owner/the owner of store)*
10) *sepatu Nike (the Nike shoes)*

Example 1-3 are noun phrases with a determiner (DET). Example 1 uses a demonstrative determiner. While in English this kind of determiner is placed before the noun, in Indonesian it is written after the noun. Example 2 dan 3 use quantity determiner that is either a number or words that describe number such as *beberapa (some/several), semua (all)*. This kind of noun phrase has the same syntax with English, a determiner is written before the noun.

Example 4 is a noun phrase with an adjective as the second word that describes the first word. Example 5 is a noun compound that we discuss in Subsection 2.3 with the syntax of "NOUN + NOUN".

Example 6-8 are noun phrases that show ownership. Example 6 uses *-ku* clitic as the possessive pronoun. Example 7-8 use "NOUN + NOUN/PROPN" syntax where the second token is the owner of the first token. While in English the ownership is marked by the *'s* clitic or using the *of* preposition, there's no such syntax in Indonesian.

Example 9-10 are noun phrases with the same syntax with Example 7-8, but with different semantics. In these phrases, the second word is not the owner of the first word, but only describes it. To differentiate between those two kinds of "NOUN + NOUN/PROPN" phrases require the knowledge of whether it is possible for the second word to 'own' the first word, that makes this task challenging.

3 Indonesian Parallel Universal Dependencies (PUD) Treebank

In this section, we discuss the treebank being revised, the UD Indonesian-PUD.

3.1 Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically grammatical annotation for dependency treebank. Initially, de Marneffe et al. (2006) designed type dependency for English that later was called Stanford Dependencies. Stanford

Dependencies scheme was designed to represent English grammatical relations between words in a sentence (de Marneffe and Manning, 2008). This representation was later adopted by de Marneffe et al. (2014) to create universal dependencies that can be applied to other languages to support cross-linguistically parsing. This new scheme was named Universal Dependencies (UD).

The first version of UD annotation guidelines was called UDv1 (Nivre et al., 2016). The recent version of the annotation guidelines is UDv2, that has tagset of 17 POS tags and has 37 dependency relations plus some dependency relation subtypes to be used by certain languages to adapt to UD.

3.2 Parallel Multilingual Treebanks

Parallel Universal Dependencies (PUD) treebanks created for *CoNLL 2017 share task for Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2018). They created parallel treebanks for 18 different languages. Each treebank consists of 1,000 sentences, in the same order. The sources of the sentences are from news domain and Wikipedia. The original language of the first 750 sentences is English, and the rest are German, French, Italian and Spanish.

3.3 Indonesian-PUD Treebank

UD Indonesian-PUD (hereinafter referred to as ID-PUD) is part of PUD. We observed that the ID-PUD has some problems, where the major flaws are its word segmentation and POS tagging.

For the word segmentation, the list of error is as follows: 1) The words with reduplication are always be split into multiple tokens; 2) Other hyphenated words are also always separated into multiple tokens; 3) The clitic *-nya* is always separated from its parent word, despite its context; 4) Many tokens that are composed of two base words such as *ketidaksesuaian (the discordance)* were separated into two tokens, while it should remain as one token.

For the POS tagging, a lot of tokens were incorrectly labeled, either a verb was labeled as a noun or a noun labeled as a verb. We suspect this problem happened because the tool determined the POS tag based on the base-word of the verb. If the base-word is a noun, then the verb is labeled as a noun, which is incorrect.

4 Adjusting UDv2 for Indonesian

This section presents our proposed annotation guidelines for the specific characteristics of Indonesian grammar.

4.1 Word Segmentation and POS Tagging as an Inseparable Task

For most word segmentation cases, all clitics should be separated from its parent word. What makes this task difficult is that for the clitic of *-nya* there are cases when it should not be separated, as explained in Subsection 2.2. For example, we have two sentences contains the word *dibukanya*:
a) ***Dibukanya*** *toko itu menimbulkan kemacetan. (The opening of the store caused a traffic jam.)*
b) *Paket itu* ***dibukanya*** *dengan hati-hati. (The package was opened (by her/him) carefully.*
For sentence (a), *dibukanya* should not be separated since it has a role as a NOUN, and for sentence (b), the token *dibukanya* should be split since it contains two syntactic token, *dibuka (was opened)* as a VERB and *-nya (him/her)* as a PRON.

To decide whether we will split a token ended with *-nya*, we proposed this general approach: 1) Split the token ended with *-nya* into two parts, the parent token and *-nya* ; 2) Determine the POS tag of the parent token; 3) Use Table 1 as the reference to solve the ambiguity by using the POS tag of tokens before or after the examined token; and 4) Finally, if the final POS tag of the examined token is a NOUN or an ADV, re-merge the parent token and *-nya*. We leave the details of this approach for future work.

Thus, the word segmentation task needs POS tags information to decide whether to split tokens ended with *-nya* or not. That's why word segmentation and POS tagging should become an inseparable task.

4.2 Adjusting Dependency Relations to Indonesian Grammar

UDv2 defined 37 dependency relation labels plus some subtypes of dependency relations to comply with special characteristics of certain languages. For current Indonesian treebanks in UD, 13 subtypes are used as shown in Table 4.

After analyzing those 13 subtypes, we proposed to retain subtypes No. 1-8 and to remove the remaining 5 subtypes. Also, we adopted 7 subtypes used by

No	Deprel	Description
1	acl:relcl	for relative clause
2	cc:preconj	for pre-conjunction
3	csubj:pass	subject clause of passive
4	nsubj:pass	subject of passive sentence
5	dep:prt	for clitic of particle
6	nmod:poss	for phrase of ownership
7	obl:tmod	noun phrase of time
8	flat:name	for named entities
9	compound:plur	for reduplicated words
10	obl:poss	for phrase of ownership
11	compound:n	for noun compound
12	compound:v	for verb compound
13	compound:a	for adjective compound

Table 4: List of subtypes used by current Indonesian treebanks in UD.

other languages and proposed the use of a new subtype. In total, we proposed the use of 16 subtypes for annotating Indonesian dependency treebank.

4.2.1 Removing five subtypes

The following is the explanation of why we propose not to use subtypes of *compound:plur*, *obl:poss*, *compound:n*, *compound:v*, and *compound:a*.

In the original ID-PUD, the reduplicated words are split into three tokens. For example, *anak-anak (children)* was split into *anak*, -, and *anak*. Subtype of *compound:plur* was created to link the third token to the first one. Since we opted not to split the reduplicated words, we no longer need this subtype.

The subtype of *obl:poss* most likely was created due to incorrect POS tagging of some nouns that labeled as verbs. For example, in ID-PUD noun phrase of *kehidupan kita (our life)* has POS tags of "VERB + PRON", while the correct POS tags should be "NOUN + DET". The correct relation between *kita (our)* to *kehidupan (life)* should be *nmod:poss*. There is no need to define *obl:poss* subtype since that case, the noun phrase with syntax of "VERB + PRON" for ownership, never exist.

UDv2 has *compound* label for noun phrases with syntax of "NOUN/PROPN + NOUN/PROPN". Since in the original ID-PUD there are noun phrases with syntax of "VERB + NOUN" such as in *bela diri (self-defense)* or "NOUN + ADJ" such as in *rumah sakit (hospital)*, a new subtype of *compound:n* was created. Since all noun phrases should have syntax shown by Table 3, we suggest to solve this problem by improving the quality of POS tagger, instead of introducing this subtype.

The subtypes of *compound:v* and *compound:a* were used for verb and adjective compound in the original ID-PUD. Table 2.3 shows that besides these two types of compound words, in Indonesian grammar there are also compound of adverb, number, determiner, and subordinating conjunction.

Because the number of compound words other than nouns is limited, we proposed that the compound words of verb, adjective, adverb, number, and determiner to be represented by only one single label. Since in English treebank ***compound:prt*** subtype was used for verb compound, we proposed the used of that label for those five compound word types in Indonesian grammar. As for the compound word of subordinating conjunction (SCONJ) that can be regarded as the function word, we proposed to use *fixed* label as suggested by UDv2 guidelines.

4.2.2 Adopting other six subtypes from treebanks of other languages

Besides adopting *compound:prt* for compound words, we also proposed the adoption of other six subtypes defined for other languages in UDv2: 1) *flat:foreign*; 2) *flat:range*; 3) *nmod:npmod*; 4) *nmod:tmod*; 5) *obl:agent*; and 6) *obl:mod*.

In UDv2 guidelines about *flat* subtype, *flat:foreign* is used to annotate a foreign phrase that cannot be given a compositional analysis. Subtype of *flat:range* was used by Ukranian PUD treebank to label the dependent of noun phrase like *"2018-2019"* or *"8 until 10"*. We considered this annotation scheme better than the current *nummod* subtype used in the original ID-PUD for this case, since *nummod* was initially designed for noun phrase with quantity determiner, such as in *5 buku (five books)*.

In ID-PUD, noun phrases with the syntax of "NOUN/PROPN + NOUN/PROPN" are labeled as a *compound*, even if the semantics of the phrase is far away from the definition of compound discussed in Subsection 2.3. We propose to use *compound* label only for noun phrases with syntax of "NOUN + NOUN". For "NOUN + PROPN" or "PROPN + NOUN" we proposed the use of *nmod:npmod* subtype instead. For example, for phrase *ibukota Indonesia*, the word *Indonesia* was given label of *nmod:npmod*. As for phrases of

"PROPN + PROPN" the *flat* label should be used as suggested by UDv2 guidelines.

Since *obl:tmod* label has been used for noun/noun phrase related to the time that describes the predicate, we propose to also use *nmod:tmod* subtype for noun/noun phrase related to time that describes the noun, such as in *laporan 2019* (the report of 2019) where 2019 (the year) describes the noun of *laporan (report)*.

In Alwi et al. (1998), it stated the passive sentence does not have an object but could have a noun/noun phrase that represents the agent. For this purpose, there are two possible labels. If it is a noun phrase with preposition we used *obl* label, but if the agent is written without preposition, *obl:agent* subtype will be used.

Subtype of *obl:mod* is initially used by French treebank for nominal adjunct of a predicate. We want to adopt this label for noun/noun phrase without preposition that describes the predicate but not the object nor the agent of the predicate. For example in the sentence *"Bunga bank naik 1%" (Bank interest rose 5%)*, the token % will be given *obl:mod* label.

4.2.3 Proposing a new subtype

We observed that some adverbs in Indonesian can be formed with the syntax of "*secara/dengan (with)* + ADJ/VERB/NOUN". Examples of such adverbs are *secara bijaksana (wisely), dengan bersemangat (excitedly), dengan setara (equally)*.

According to UDv2, since *secara* or *dengan* are the prepositions, their POS tag is ADP if followed by a noun phrase or SCONJ if followed by a clause. In syntactic parsing, the token with ADP tag will be labeled with *case* label and the token with SCONJ tag with *mark* label.

On the other hand, the syntax of "*secara/dengan* + ADJ/VERB/NOUN" in forming an adverb in Indonesian grammar needs special treatment. We proposed a new label named ***case:adv*** for *secara/dengan* and for the ADJ/VERB/NOUN following them, its POS tag need to be changed to ADV so that we can label them as *advmod*. It will be the responsibility of the POS tagger to identify this kind of adverb in sentences and to modify the POS tag of the related words.

Figure 1 shows an example of how a dependency tree has changed. A reduplicated word *saudara-saudara (folks)* was split into three tokens in the original treebank, but remain as one token in the revised treebank. Additionally, we also revised the POS tag of word *yang (that)* and changed the subject of this sentence.

5 Revising the Indonesian PUD Treebank

In this section, we present the annotation procedure and the statistics of the revised treebank.

5.1 Annotation Procedure

The revision was done in 2 stages: 1) Revising the word segmentation and POS tags; 2) Revising the dependencies. Both stages were done by two annotators, with the background of computer science and Indonesian linguistics. The total time for learning the UDv2 annotation guidelines, proposing the adjustment for Indonesian grammar and conducting the revision of ID-PUD treebank was six months.

For each stage, the revising was done in two phases: the learning phase and the revision phase. On the learning phase, each annotator was given 50 first sentences of ID-PUD to be analyzed. On the meeting, the annotators discussed what should be done in revising the treebank by referring to UDv2 guidelines and references of Indonesian grammar. After both annotators agree on all issues, the revision phase was started. The process was done iteratively. If there was a new case found in the revision phase, the annotators were back to the learning phase and update the guidelines. After that, the revision phase was resumed.

5.2 Statistics of the Revised Treebank

Table 5 shows the comparison of token distribution in the original and revised treebanks, as the effect of changes in the word segmentation process. Since in the revised UD-PUD, we did not split the reduplicated words and a lot of other hyphenated words, the number of tokens is smaller compared to the original ID-PUD. Likewise, the average number of tokens in the sentence becomes smaller, and the number of unique tokens increased.

Table 6 shows the comparison of UPOS distribution in the original and revised treebank. It shows that major revision had been done in the POS tagging process. For example, there is no SCONJ and

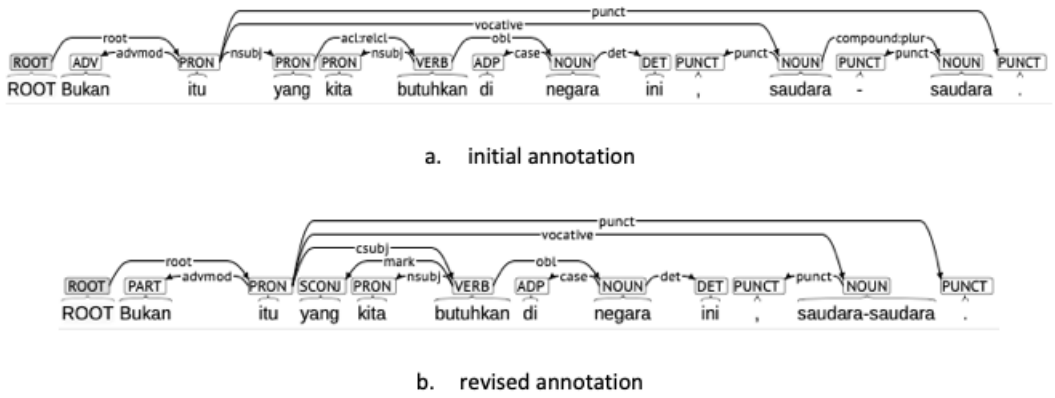

Figure 1: The initial and revised annotation for a sentence of *"Bukan itu yang kita butuhkan di negara ini, saudara-saudara." (That's not what we need in our country, folks.)*

Description	Original	Revised
Number of tokens	19,900	19,401
Avg number of tokens in sentence	19.9	19.4
Number of unique tokens	4,692	4,732

Table 5: Statistics of number of tokens.

INTJ in the original treebank, but in the revised one we have 487 occurrences of SCONJ and 4 occurrences of INTJ.

UPOS	Ori	Rev	UPOS	Ori	Rev
ADJ	1358	962	PART	57	276
ADP	2832	1901	PRON	989	1049
ADV	1049	623	PROPN	1456	2217
AUX	211	424	PUNCT	2579	2384
CCONJ	612	595	SCONJ	0	487
DET	522	940	SYM	37	39
INTJ	0	4	VERB	1965	2359
NOUN	5578	4618	X	86	17
NUM	569	506			

Table 6: The UPOS distribution.

As for the dependency relation labels, in the new treebank, we only use 32 of 37 UDv2's main labels and 15 of 16 subtypes described in Subsection 4.2. UDv2's main labels that were not used are *clf, dep, expl, list*, and *reparandum*, while the subtype not used is *flat:name*. In total, the revised ID-PUD used 47 labels.

We decided not to use *flat:name* for names and used *flat* for all proper noun instead since it's still not clear for us which names are suitable for *flat:name*. Once we know how to differentiate between *flat* and *flat:names* we will revise the treebank.

6 Experiments and Results

To evaluate the quality of the revised ID-PUD, we built the Indonesian parser model using Parsito (UDPipe) that built by Straka et al. (2015). Parsito is a transition-based parser that utilized neural network classifier for prediction and requires no feature engineering. We used this parser with default parameter.

Accuracy was evaluated using the ten-fold cross-validation method. The performance measurements used are UAS (Unlabeled Attachment Scores) and LAS (Labeled Attachment Score) (Kübler et al., 2009). Table 7 shows the comparison of accuracy between the original and revised ID-PUD treebank.

Treebank	UAS	LAS
Original	73.32%	65.98%
Revised	**83.33%**	**79.39%**

Table 7: Experiment results.

The result shows that the model built by our revised treebank has higher UAS and LAS than the original one, with a margin of 10% for UAS and around 13% for LAS. It shows that the revised treebank has better consistency in annotation so that the learning algorithms can learn the pattern better than when using the original one.

To find out which labels had achieved good accuracy, we used MaltEval (Nilsson and Nivre, 2008) to compute the F1-score of 47 labels used in the revised treebank and shown the result in Figure 2.

We had a hypothesis that there is a correlation between F1-score and the number of occurrences of

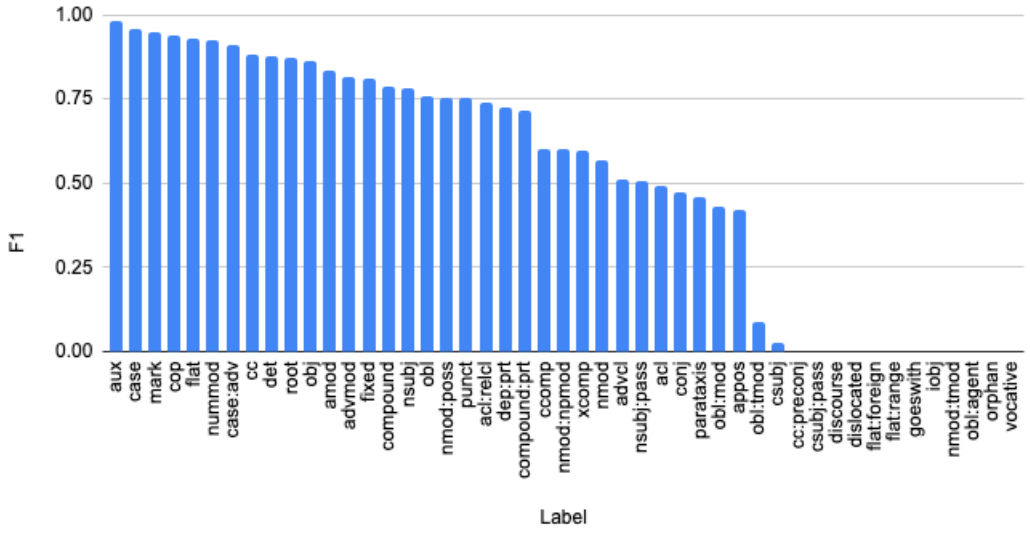

Figure 2: The F1-score of dependency labels in the revised ID-PUD.

each label, but that isn't true because the correlation coefficient is only 56%. For example, *case:adv* that occurs only 57 times has F1-score of 91%, while *conj* which occurs 666 times, has F1-score of 47%. We suggest that the low F-score was caused due to the lack of consistent patterns for those labels.

However, for those labels that occurs only 10 times or less, we believed that the F1-score can be improved by increasing the size of the treebank and adding more examples with those labels.

To improve the accuracy of the Indonesian parser model, we have three suggestions: 1) to revisit the choices of the dependency labels so that each label was designed with distinct characteristics; 2) to revisit the annotation whether the rules had been applied consistently; and 3) to employ additional morphology features that have not been added to this revised ID-PUD treebank.

7 Conclusions and Future Work

We proposed a revision to an existing dependency treebank in Indonesian, named ID-PUD that consists of 1,000 sentences. The annotation was done manually, refers to UDv2 annotation guidelines and the references of Indonesian grammar. Besides, we also proposed how to conduct word segmentation and POS tagging for Indonesian sentences, especially related to the handling of *-nya*. Some changes in dependency labels for Indonesian dependency treebank are also proposed, resulting in the 16 subtypes to adjust to Indonesian grammar rules.

To evaluate the quality of the new treebank, we used Parsito (UDPipe) parser to build the parser model using 10-fold cross-validation method. The results show that the model built using the revised treebank has a higher UAS and LAS with the margin of more than 10% than the original ID-PUD. This shows that the new treebank has a better label consistency compared to the original one.

This manual revision of ID-PUD took so much time and efforts. In future work, we want to build tools to automate the word segmentation and POS tagging described in this paper. Besides that, adding morphology features needs to be done so that this treebank has the same attributes with other parallel treebanks in PUD.

Acknowledgments

This work was supported by the research grant of PTUPT (Penelitian Terapan Unggulan Perguruan Tinggi) No. NKB-1693/UN2.R3.1/HKP.05.00/2019 from the Ministry of Research and Technology, Republic of Indonesia.

References

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 1998. *Tata Bahasa Baku Bahasa Indonesia*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation - CrossParser '08*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford Dependencies : A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

John A Hawkins. 1990. A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, 21(2):223–261.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool.

Septina Dian Larasati. 2012. Handling Indonesian Clitics: A Dataset Comparison for an Indonesian-English Statistical Machine Translation System. In *26th Pacific Asia Conference on Language, Information and Computation*, pages 146–152.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbachbrundage, Yoav Goldberg, Dipanjan Das, Kusman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Jens Nilsson and Joakim Nivre. 2008. MaltEval : An Evaluation and Visualization Tool for Dependency Parsing. In *LREC 2008*, pages 161–166.

Joakim Nivre, Marie-catherine De Marneffe Filip, Ginter Yoav, Jan Haji, D Manning Ryan, Mcdonald Slav, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. pages 1659–1666.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. A&U Academic.

Milan Straka, Jan Hajič, Jana Straková, and Jr. Jan Hajič. 2015. Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle. In *14th International Workshop on Treebanks and Linguistics Theories (TLT 14)*, pages 208–220.

Daniel Zeman, Jan Haji, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.