A Dependency Annotation Scheme for Indonesian

Budi Irmawati* Hiroyuki Shindo Yuji Matsumoto
Nara Institute of Science and Technology
{budi-i, shindo, matsu}@is.naist.jp

Abstract

This paper proposes a dependency annotation scheme for Indonesian. Our scheme basically follows the Stanford typed dependency annotation, however, we propose some adaptations for morphologically-rich phenomena in Indonesian, such as affixations, ellipses, and non-verb clauses. This paper presents those phenomena and describes how the scheme accommodates the phenomena in deciding dependency relations between two words. Evaluation of this scheme by training it on the MST parser is also reported.

1 Introduction

Indonesian has subject-verb-object (SVO) construction but has relatively free word order in terms of noun phrases and verb phrases. Although in general they are postmodifier phrases, most position of word modifier is fixed, relative to the word that is modified by the modifier; either before or after the modifee.

Kübler et al. [4] stated that dependency grammar is better than phrase structure grammar for languages with free word order. Moreover, Rozovskaya et al. [7] reported some error correction tasks and showed that dependency relations improved the tasks. These dependency relations also worked well in our experiment of detection error types and error positions of sentences written by second language (L2) learners, in which the POS tags and language models are not strong enough to detect preposition, verb, and noun error types. Therefore, in the absence of a dependency parser, we defined a basic annotation scheme to build a dependency parser model.

This paper is arranged as follows. In Section 2, we summarize some previous studies on the dependency resources of Indonesian and a dependency annotation scheme. Section 3 describes the annotation scheme followed by the evaluation method in Section 4. We explain the results in Section 5, then, we conclude our work and describe the future plans in the last section.

*on leave from University of Mataram, Indonesia

2 Related work

Currently, dependency resources for Indonesian language is not publicly available. Even though there are two works on dependency parsing [2, 3], they used their internal resources. Moreover, Green et al. [2] only described, in general, their unlabeled head assignment in seven rules, which the trees have the verb as the root of the sentence. Our annotation is one step further by considering the non-verb clauses.

McDonald et al. [6] defined universal dependency annotation for German, English, Swedish, Spanish, French, and Korean. However, Indonesian was not covered by this annotation.

3 The annotation scheme

The annotation scheme follows the Stanford typed dependency (SD) manual [1] adapted to the language phenomena of Indonesian that affect dependency annotation. After the data preprocessing, we briefly discuss the "phenomena" based on [8] in Subsection 3.2 followed by the adaptations and added labels in Subsections 3.3 and 3.4.

3.1 Data Preprocessing

The dependency annotation scheme is applied on 650 error-corrected learner sentence pairs. Both learner sentences and the corresponding corrections were annotated by a native speaker, but this experiment only used the corrected sentences to evaluate the annotation scheme.

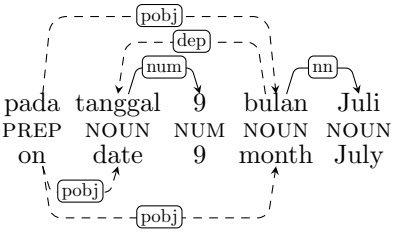
Most of the sentences are declarative sentences and some sentences are imperative sentences. All the sentences have been POS tagged[1], split from clitic, and chunked automatically. Clitics are separated from main words because they can be a subject, an object, a possessive pronoun, or a determiner.

To decrease the complexity of the phrases, the sentences were chunked using a simple finite automaton (FA) to decide the beginning and the end of the chunk. The FA is not explained here because of limited space, but the example below shows a complex

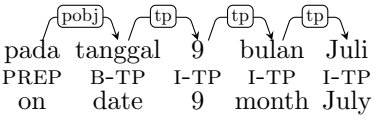
[1]sentences is tagged using Morphind (http://septinalarasati.com/work/morphind/)

time phrase converted into the simpler one. In example (1a), the dashed lines above and below the sentence are two possible relations for this time phrase before chunking, while (1b) shows simpler relation after chunking.

(1) a.



b.



'on 9 July'

3.2 Phenomena in Indonesian

This section describes some phenomena that are not covered in SD annotation. Some adjectives can be used as an adverb while a clitic has different functions depending on the word it precedes or follows. Ellipsis is an omission of a word from a sentence if its presence is unnecessary.

To prevent the relations to be affected by an ellipsis, we take care of the ellipsis of verbs and linking verbs; and also non-verb heads because a verb is usually needed to become the head of a clause. This subsection describes adverbs, adjectives, clitics, determiners, non-verb clauses, and copulas.

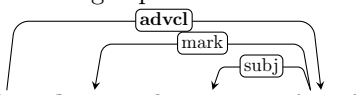
- **Adverb**. Some adverbs can be formed from adjectives. Some of this adjectives should be preceded by *dengan* 'with'.
- **Adjective**. Some adjectives need to be preceded by *yang* 'which is' to keep the naturalness of the sentence.
- **Clitic**. Indonesian uses front-clitics (*ku-* [1SG] and *kau-* [2SG]) as a subject of a clause while end-clitics (*-ku* [1SG], *-mu* [2SG], and *-nya* [3SG]) as a direct object, an indirect object, an object of preposition, a possessive pronoun, or a determiner.
- **Determiner**. Determiners are used to point which object is referred to by the writer or to show the cardinality of the object but Indonesian optionally uses a determiner for a singular object. Some determiners precede the noun they modify.
- **Non-verb clauses**. The head of the clause is not always a verb. It can be a noun or an adjective.
- **Copula** corresponds to a linking verb in grammar of other languages. A copula occurs optionally between the subject and the predicate in non verbal clauses. A copula is not obligatory, except where either the subject or the predicate is long.

3.3 Adaptation to Stanford typed dependency annotation

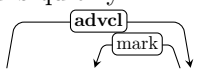
We propose some adaptations for Indonesian dependency annotation scheme because the language phenomena described in 3.2 are not covered in the SD annotation.

- ***adapt-1***, **advcl** labels an adverbial clause that modifies a VP as in (2a). Since the adjective preceded by *dengan* 'with' is used as an adverb and modifies a VP, the **advcl** labels this adjective as in (2b).

(2) a. Ia berlari dengan kecepatan
3SG ACT.run PREP.with NOUN.speed
tinggi
ADJ.high
'S/he runs in high speed'

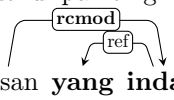


b. Ia berlari dengan cepat
3SG ACT.run PREP.with ADJ.cepat
'S/he runs quickly'



- ***adapt-2***, **rcmod** labels a relative clause that modifies an NP. Since an adjective preceded by *yang* 'which is' also modifies an NP, **rcmod** labels this adjective as in (3).

(3) lukisan yang indah
NOUN.painting which is ADJ.beautiful
'beautiful painting'



- ***adapt-3***, **mark** introduces an adverbial phrase. Since the adjective in ***adapt-1*** comes after *dengan* 'with' is used as an adverb, **mark** labels *dengan* as in (2b).

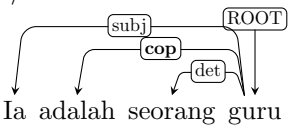
- ***adapt-4***, **ref** introduces a relative clause that modifies an NP. *yang* 'which is' precedes the adjective that modifies an NP, so the **ref** labels *yang* as in (3).

- ***adapt-5***, labels all front-clitics as **subj** and labels end-clitic as **dobj** if the clitic follows a transitive verb; as **pobj** if the clitic follows a preposition; or as **iobj** if the clitic follows a ditransitive verb. ***adapt-5*** also labels the clitic as **det** if the clitic is used as a determiner or **poss** if the clitic is used as a possessive pronoun.

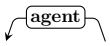
- ***adapt-6***, adds **tp** to label the words inside the time chunk.

- ***adapt-7***, ***cop*** labels a copula. Copula is set as a modifier of a noun as in (4), so if the copula is omitted from a sentence, it will not change other relations in the sentence.

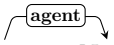
(4) Ia adalah seorang guru
3SG COP.is DET.a NOUN.teacher
'S/he is a teacher'



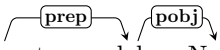
- ***adapt-8***, **agent** labels an actor in passive type-II as in (5a) because in passive type-I, **agent** labels the actor only if the actor is not preceded by the preposition *oleh* 'by' as in (5b). If the actor is preceded by *oleh*, it labels as **pobj** as in (5c). The passive voices (type-I and type-II) are described in detail in [8].

(5) a.


saya Nani jemput
1SG Nani ACT.pick.up

b.


saya dijemput Nani
1SG PASS.pick.up Nani

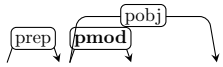
c.


saya dijemput oleh Nani
1SG PASS.pick.up PREP.by Nani

'I am picked up by Nani'

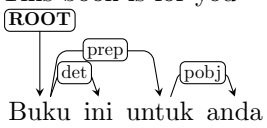
- ***adapt-9***, instead of as an independent word, a compound preposition is treated as a chunk. Therefore, ***adapt-9*** adds **pmod** to label the words inside the compound preposition such as *samping* for word *di samping* 'beside' as in 6.

(6) Ia berdiri di.samping saya
3SG ACT.stand PREP.beside 1SG
'S/he stands beside me'



- ***adapt-10***, In non-verbal clauses, the head of noun clauses, adjective clauses, or quantity clauses is the noun or the adjective, but the noun in prepositional clauses cannot be chose as the head of the clauses. In prepositional clause case in (7), the topic is chosen as the head.

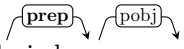
(7) Buku ini untuk anda
NOUN.book DET.this PREP.for 2SG
'This book is for you'



3.4 Added Label

After evaluating our adaptation scheme, we found that MST parser assigned incorrect heads for the adjectives preceded by *yang* or *dengan* because *yang* can also be used as a pronoun while *dengan* is also a preposition. For these reasons, we added two labels and assigned them to the adjectives of ***adapt-1*** and ***adapt-2***. The added labels are described below.

- ***addedLabel-1***, **padv**, "adverb with preposition", replaces the **advcl** label in ***adapt-1*** of (2b) because sometimes parser set the relation in (2b) with ***prep*** as in (8).

(8)


- ***addedLabel-2***, **rpmod**, "modifier with relative pronoun", replaces the **rcmod** label as (3) of ***adapt-2*** to differentiate the adjective preceded by *yang* with an adjective that directly modifies a noun.

4 Evaluation

To check the consistency of the annotation scheme and the correctness of the annotation, we run MST parser [5] on the sentences annotated in three annotation schemes. The schemes are (1) *adaptOnly*: scheme with adaptions only without chunking; (2) *adapt+addedLabel*: scheme with adaptions and added-label but without chunking; and (3) *adapt+addedLabel+chunk*: scheme with adaptions and added-label with chunking. We split the sentences into 550 training sentences and 94 test sentences.

We evaluated the accuracy based on the complexity of the sentences: *NoClause*, *Clauses*, *NoVerbRoot+Clauses*, and *Subjectless* with the proportion of the sentences shown in Table 1. The *NoClause* are simple sentences that only have subject, verb, and might have an object and prepositional phrases; the *Clauses* are sentences with one or more clauses including prepositional phrases; the *NoVerbRoot+Clauses* are sentences that their root is not a verb but they might have clauses; and the *Subjectless* are sentences that do not have subject such as imperative sentences.

5 Experimental Results

Tables 2, 3, and 4 show the unlabeled accuracy score (UAS) and labeled accuracy score (LAS) of the three

Table 1: Sentence categories based on the presence of clause

Complexities	#Sent	#Token
All	94	929
NoClause	27	202
Clauses	35	298
NoVerbRoot+Clauses	28	210
Subjectless	4	19

Table 2: *adaptOnly* scheme

Complexities	Accuracy		Complete	
	UAS	LAS	UAS	LAS
All	0.646	0.576	0.217	0.130
NoClause	0.797	0.703	0.370	0.222
Clauses	0.638	0.581	0.104	0.125
NonVerbRoot+Clauses	0.581	0.5	0.179	0.063
Subjectless	0.737	0.632	0.25	0.0

different scenarios. Table 3 shows that added the labels increased the dependency accuracy by 0.13 point compared to *adaptOnly* scheme and Table 4 shows that sentences previously chunked have better accuracy by 0.02 point compared to the sentences that were not chunked, both on UAS and LAS.

The head of adverbial clauses and relative clauses are mostly a verb, which are labeled with **advcl** or **rcmod** depending on ***adapt-1*** or ***adapt-2***. After ***addedLabel-1***, the preposition *dengan* 'with', which introduces an adverb to modify a verb, is not labeled as **prep** and after ***addedLabel-2***, the **rcmod** relation becomes more consistent because the adjectives are labeled with **rpmod** if they are used as an adjective preceded by *yang* 'which'.

However, parser sometimes mistakenly labels a clitic as either a determiner or a possessive pronoun because both clitics follow a noun. The parser also mistakenly labels a clitic as **iobj** for a direct object because the training data contain very few examples of a clitic labeled as **dobj**.

6 Conclusion

This experiment shows that updating head and label assignment and adding new labels based on the phenomena of the language improve the parser accuracy. Moreover, for some complex phrases, chunking also eased the parser in assigning the in-chunk relations.

We used this annotation scheme to generate a parser model for our grammatical error-correction task. For future direction we need to improve the parsing pre-process and polish up the scheme to reduce parser errors and handle interrogative sentences

Table 3: *adapt+addedLabel* scheme

Complexities	Accuracy		Complete	
	UAS	LAS	UAS	LAS
All	0.788	0.707	0.357	0.214
NoClause	0.874	0.8	0.516	0.370
Clauses	0.723	0.701	0.191	0.063
NonVerbRoot+Clauses	0.725	0.642	0.357	0.214
Subjectless	0.842	0.842	0.5	0.5

Table 4: *adapt+addedLabel+chunk* scheme

Complexities	Accuracy		Complete	
	UAS	LAS	UAS	LAS
All	0.812	0.731	0.394	0.214
NoClause	0.944	0.877	0.733	0.467
Clauses	0.825	0.736	0.345	0.087
NonVerbRoot+Clauses	0.841	0.754	0.167	0.167
Subjectless	0.842	0.789	0.25	0.5

and long distance clauses.

Acknowledgments

This study is supported in part by the DGHE, Republic of Indonesia under BPPLN Scholarship Batch 7 fiscal year 2012-2015.

References

[1] M. De Marneffe and C. D. Manning. Stanford Typed Dependencies Manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

[2] N. Green, S. D. Larasati, and Z. Žabokrtský. Indonesian Dependency Treebank: Annotation and Parsing. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, Bali,Indonesia, 2012.

[3] M. Kamayani and A. Purwarianti. Dependency Parsing for Indonesian. In *International Conference on Electrical Engineering and Informatics, ICEEI*, pages 1–5. IEEE, 2011.

[4] S. Kübler, R. T. McDonald, and J. Nivre. *Dependency Parsing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2009.

[5] R. McDonald, K. Lerman, and F. Pereira. Multilingual Dependency Analysis with a Two-stage Discriminative Parser. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 216–220, 2006.

[6] R. McDonald, J. Nivre, Y. Quirmbach-brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Tckstrm, C. Bedini, N. Bertomeu, and C. J. Lee. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL'13*, 2013.

[7] A. Rozovskaya, D. Roth, and V. Srikumar. Correcting Grammatical Verb Errors. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 358–367, Gothenburg, Sweden, 2014.

[8] J. N. Sneddon, A. Adelaar, D. D. N., and M. C. Ewing. *Indonesian: A Comprehensive Grammar*. Routledge, Australia, 2010.