INARTE: An Indonesian Dataset for Recognition Textual Entailment

Abdiansah Abdiansah
Faculty of Computer Science
Universitas Sriwijaya
South-Sumatera, Indonesia
abdiansah@unsri.ac.id

Azhari Azhari
Department of Computer Science
Universitas Gadjah Mada
Yogyakarta, Indonesia
arisn@ugm.ac.id

Anny Kartika Sari
Department of Computer Science
Universitas Gadjah Mada
Yogyakarta, Indonesia
a_kartiksari@ugm.ac.id

***Abstract*— Recognition Textual Entailment (RTE) try to solve variability problem that commonly encountered in natural language-based systems. The basic idea is to detect whether the meaning of a text can be inferred by another text. The need dataset in language other than English is necessary to accelerate research development in RTE. We created RTE dataset for Indonesian by retrieval text from Web and generate text-hypothesis pairs as many as possible. The subset technique is used to decide whether Text (T) entails Hypothesis (H). The initial data used 400 question-answer pairs obtained 1,577 entailment pairs, where 481 entailment pairs obtained from the accuracy above 50%.**

Keywords—RTE, entailment, variability, text, hypothesis

I. INTRODUCTION

In recent years, a number of Natural Language Processing (NLP) researchers have developed and participated in the task of Recognition Textual Entailment (RTE) [1]. RTE has been proposed as a generic task that captures major semantic inference needs across different applications, such as Question Answering System (QAS), Information Extraction (IE), Information Retrieval (IR), Machine Translation (MT), Summarization and Paraphrasing [2]. Also, the task encapsulates Natural Language Understanding (NLU) capabilities within a very simple interface: "recognizing when the meaning of a text snippet is contained in the meaning of a second piece of text". Since RTE task has been defined recently, there exist only few dataset for training and testing RTE systems, and none of them are in Indonesian. Thus, we planned the development of INARTE, a dataset for traning and testing RTE systems in Indonesian.

We built RTE dataset following Penas et al. [3], but we do not fully implement because there are several different cases, such as type of data source, system evaluation, and others. The data sources used by [3] comes from real Spanish QA systems. They take three kinds of data from the systems, namely (1) questions, which is a user query to the system; (2) documents, set of text in which the answers is located; and (3) answers, the system result. Each question-answer pairs are manually checked by humans to judge that the pairs have entailment or not. The approach of Penas et al. [3] needs to be modified if it applied to the Indonesian language. This is due to the low availability of Indonesian QA systems causing us difficulties to collecting data. To solve the problem, we created a dataset containing question-answer pairs which is labeled TRUE if the answer is correct and FALSE if it is incorrect. The question-answer format is simple as follows {"question", "answer"} and the example of question-answer (labeled TRUE) as follows {"Who is the first general of VOC?", "Pieter Both"}. Furthermore, the question-answer pairs was converted to affirmative form so that it would be {"Pieter Both is the first president of VOC"} and called a hypothesis. The hypothesis is considered as keyword and sent to the search engine to retrieve set of relevant sentences from Web. The relevant sentences obtained are called text. Each hypothesis will be paired to their relevant sentences. After that we evaluated the text-hypothesis pairs whether they contain entailment or not.

This paper organized as follows, Section 2 describes the works related to our work, notably Penas et al. [3] which is our main reference. Section 3 describes the development stages of INARTE from data collection stage to final stage. Section 4 explains the evaluation specification of INARTE. Section 5 shows and discusses the results of experiment, and the last Section is conclusion and future work.

II. RELATED WORK

The topic of textual entailment originally proposed by Dagan and Glickman [4] and subsequently established through the series of benchmarks known as the PASCAL Recognising Textual Entailment (RTE) Challenges [5]. These challenges provided researchers with concrete datasets on which they could evaluate their approaches, as well as a forum for presenting, discussing, and comparing their results [2]. Unfortunately, the target language of the datasets only focus on English and need extra efforts for other languages, especially in low-resources language (ex. Indonesian). Some researchers have developed RTE datasets in their native languages including Spanish [3], Arabic [6], German [7], and Czechoslovakia [8]. But in Indonesian language, to our knowledge it has not been done.

Our work is similar to Penas et al. [3] because we get inspiration from them. But there are some fairly basic differences, namely: (1) They obtains question-answer pairs and documents from real QAS. Whereas we use datasets containing question-answer pairs and documents retrieved from Web; (2) They focuses on assigning entailment values of text-hypothesis pairs by annotators, since their data are complete (questions, answers and documents). While we focus on searching documents on Web with queries from question-answer pairs; and (3) They does not perform language

978-1-5386-5813-9/18/$31.00 ©2018 IEEE

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on January 29,2023 at 13:55:10 UTC from IEEE Xplore. Restrictions apply.

processing operations (parsing, tokenization, etc.), they only analyze the QAS's output. Whereas we need text processing tools to process raw data from Web. Although we used different methodology but the goal is same which is to find as many as possible text-hypothesis pairs.

III. DEVELOPMENT OF INARTE

The work of INARTE is still in progress. Some language processing modules that developed are still being refined. Figure 1 shows the general stages of INARTE's development. In the next sub-section we will explain the detail of each stages.

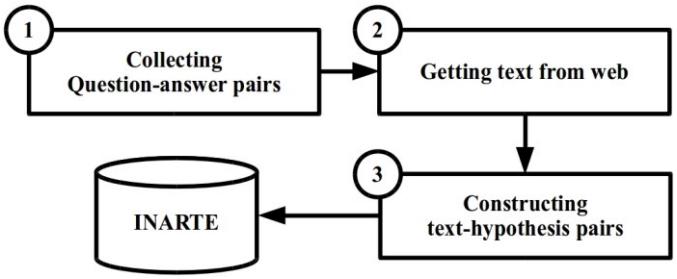

Figure 1. General of INARTE's development stages

A. Collecting question-answer pairs

The first step is collect question-answer pairs in Indonesian National History field. The reason why we chose this field is because an information related to the history is generally contains facts so that it ease to verification the truth. The questions and answers are taken from several sources including Wikipedia and history textbooks. After that the question-answer pairs are verified by experts. To facilitate the evaluation, we split the question-answer pairs into two datasets namely DS-X-R contains correct answer (TRUE) and DS-X-W contains incorrect answer (FALSE). The letter 'X' is the number of questions-answers in the dataset, for example the DS-100-R means that there are 100 questions with correct answer. We have collected 400 question-answer pairs so that we have DS-400-R and DS-400-W dataset. Both datasets will be procced and evaluated separatelly.

B. Getting text from Web

Question-answer pairs which has been collected (No. 1 in Figure 2) will be converted into an affirmative sentence by AHG or Automatic Hypothesis Generation (No. 2 in Figure 2). AHG uses rule-based technique to detecting question words and forming sentence patterns. For example, a question that begins with the phrase of "Siapa nama" (Who's name) will be applied a rule as follows:

1. Eliminated phrase of "*Siapa nama*" (Who's name)
2. Applied pattern:
   - [**jawaban** + adalah + **pertanyaan**] (answer + is + question)
3. Example:
   - DS-400-R → {*question*, *answer*}
   - DS-400-R → {*Siapa nama jenderal VOC pertama?*, *Pieter Both*}
   - AHG → {*jenderal voc pertama*, *Pieter Both*}
   - AHG → {*Pieter Both adalah jenderal voc pertama*}

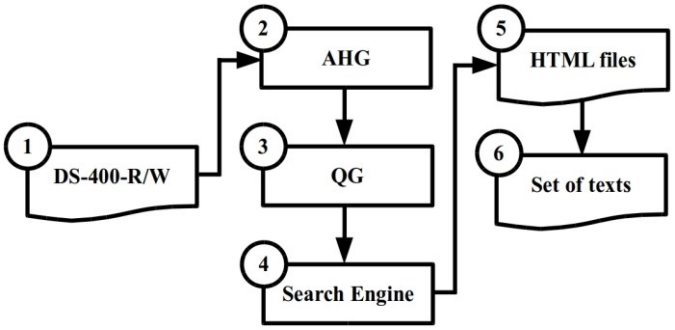

Figure 2. Steps of getting text from Web

The affirmative sentences (or hypothesis) are generated by AHG will be used as a query. Our prior experiments showed that using a single query would yield a small amount of text. To resolve this issue, we made additional queries by gradually reducing the query information. The process is performed by QG or Query Generation (No. 3 in Figure 2). Here is an example of QG results:

- **Pieter Both jenderal VOC pertama**
- Pieter Both jenderal VOC
- Pieter Both jenderal
- Pieter Both

Furthermore, Search Engine (No. 4 in Figure 2) will process the queries to get HTML files (No. 5 in Figure 2). Technically, the search engine gives a single page containing set of links that relevant to the query (10 links per page). Afterthat, each contents of link will be extracted to produce a set of text (No. 6 in Figure 2). The difficult task in this part is to cleaning up the unstructured text (html source-code) into structured text.

C. Constructing text-hypothesis pairs

After obtaining a set of text from the previous process, the next step is Sentence Tokenization (No. 2 in Figure 3). We use Punkt technique [9] to tokenized the sentences which the results is better than dot detection techniques. This technique can distinguish the dot as end of sentences or not (eg, Dr., Prof.). Furthermore, Sentence Reduction (No. 3 in Figure 3) aims to reduce long sentences without losing their original meaning. The reduction is expected to eliminate words or phrases that are less relevant to the hypothesis. Following is example of sentence reduction:

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on January 25,2023 at 13:55:10 UTC from IEEE Xplore. Restrictions apply.

- **Hypothesis** → "*Pieter Both adalah jenderal VOC pertama*"
- **Tokenization** → "*akhirnya pada tahun 1609 pieter both menjadi gubernur jenderal voc di hindia belanda (indonesia)*"
- **Sentence reduction** → "*akhirnya pada tahun 1609 pieter both menjadi gubernur jenderal pertama voc di hindia belanda (indonesia)*"
- **Sentence reduction** → "*pieter both menjadi gubernur jenderal pertama voc*"

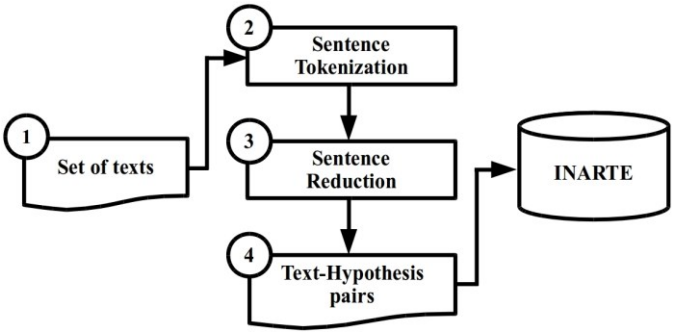

Figure 3. Steps of constructing text-hypothesis pairs

TABLE I. EXAMPLES OF TEXT-HYPOTHESIS PAIRS

Text	Hypothesis
voc memperoleh izin di banten untuk mendirikan kantor perwakilan dan pada 1610 pieter both diangkat menjadi gubernur jenderal voc pertama	pieter Both adalah jenderal voc pertama
voc mengangkat seorang pemimpin dengan pangkat gubernur jenderal gubernur jenderal voc yang pertama adalah pieter	pieter Both adalah jenderal voc pertama
soekarno lahir di blitar jawa timur 6 juni 1901 wafat di jakarta 21 juni 1970 pada umur 69 tahun adalah presiden indonesia pertama	soekarno adalah presiden indonesia pertama
presiden soekarno bung karno inilah sosok pejantan tangguh indonesia lelaki yang tampak jelas wibawa dan kharismanya ini menjabat sebagai presiden pertama	soekarno adalah presiden indonesia pertama

Here is the sentence reduction algorithm:

1. Scanning word one by one in the sentence (text), if the word is exist in the hypothesis, then mark it as left-boundary.
2. Continue the process until all words in hypothesis are also found in the text. The last word of hypothesis is marked as right-boundary.
3. Subset a text from left-boundary to right-boundary.
4. Remove all punctuation

After all sentences are reduced, the final step is to make text-hypothesis pairs (No. 4 in Figure 3) by merge the hypothesis and text which have an entailment. The relation of hypothesis and text is 1-to-n, means that one hypothesis can have many texts. Table 1 shows an example of text-hypothesis pairs. The entire pairs of text-hypothesis then we called as INARTE dataset. This dataset can be used as test data for Indonesian RTE systems to detect whether text (T) entails hypothesis (H). Test results can be compared with manual annotations performed by humans.

IV. EVALUATING INARTE

We developed TES (Textual Entailment System) to evaluate INARTE. TES uses subset to determine the entailment pair. In Eq. (1) it can be seen that T entails H is TRUE if H subset T, other than FALSE. This technique is used because easy to implement and gives standard results. It can be used as baseline for further testing. Table 2 shows examples of TES's result for text-hypothesis pairs valued TRUE and FALSE.

$$F(T,H) = \begin{cases} TRUE; H \subset T \\ FALSE; else \end{cases} \quad (1)$$

TABLE II. EXAMPLES OF ENTAILMENT

Text	Hypothesis	Entailment
voc memperoleh izin di banten untuk mendirikan kantor perwakilan dan pada 1610 pieter both diangkat menjadi gubernur jenderal voc pertama	pieter Both adalah jenderal voc pertama	TRUE
soeharto di kenal sebagai hanya satu presiden di indonesia yang mempunyai masa jabatan terlama yakni sekitaran 32 th & soeharto adalah presiden pertama indonesia	pieter Both adalah jenderal voc pertama	FALSE

We evaluated two datasets namely DS-400-R and DS-400-W. Separation is done to simplify the analysis of evaluation results. In DS-400-R is expected to generate entailment pairs as many as possible. Whereas, in DS-400-W is expected to generate entailment pairs as few as possible. Finally, we measure the accuracy of entailment pairs using Eq. (2).

$$Accuracy\ (\alpha) = \frac{\sum_{i=0}^{n} entailment-pairs}{\sum_{i=0}^{n} TH-pairs} \quad (2)$$

V. EXPERIMENTAL RESULT AND DISCUSSION

We have conducted three experiments to find out information that related to (1) Number of hypothesis containing sentences; (2) Number of hypothesis containing entailment pairs; and (3) Number of hypothesis based on accuracy filter. The initial dataset DS-400-R and DS-400-W are respectively question-answer pairs with true and false values. In Table 1, for DS-400-R it appears that the number of hypothesis containing sentences is 375 of 400 hypothesis or 93.75%. It is means that

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on January 25,2023 at 13:55:10 UTC from IEEE Xplore. Restrictions apply.

data retrieval from Web gives good results because is above 90%. Empty hypothesis will be removed from dataset. The number of sentences of DS-400-R is more than DS-400-W because it is accordance with general knowledge that the fact information more easily found on Web than not fact.

TABLE III. NUMBER OF HYPOTHESIS CONTAINS SENTENCES

A	B	C	D	E
DS-400-R	400	16.188	375 (93,75%)	25 (6,25%)
DS-400-W	400	14.351	367 (91,75%)	33 (8,24%)

[A.] Dataset name; [B.] Number of hypothesis; [C.] Number of sentences; [D.] Number of hypothesis contains sentences; [E.] Number of empty hypothesis

In Table 4, it can be seen that the dataset used is just hypothesis containing sentences. Therefore, the number of empty hypothesis (see Table 1) will be omitted. The DS-375-R dataset detected 213 hypothesis with entailment pairs (text-hypothesis TRUE) so that the total of entailment pairs is 1.577. The percentage of entailment pairs for DS-375-R and DS-367-W are respectively larger and smaller than 50% so it is considered feasible to serve as baseline testing. As mentioned in Section 4, DS-367-W is expected to provide fewer entailment pairs because the entailment pairs resulting from non-facts tend to be false positives.

TABLE IV. NUMBER OF HYPOTHESIS CONTAINS ENTAILMENT

A	B	C	D	E
DS-375-R	375	1.577	213 (56,80%)	162 (43,20%)
DS-367-W	367	299	74 (20,16%)	293 (79,83%)

[A.] Dataset name; [B.] Number of hypothesis; [C.] Number of entailment pairs; [D.] Number of hypothesis contains entailment pairs; [E.] Number of hypothesis not contains entailment pairs

Furthermore, in Table 5 and Table 6 it contains experimental results for both types of datasets. The calculation results for number of sentences and number of entailment pairs are based on the number of hypothesis. Total accuracy is obtained from summation of all hypothesis accuracy or (α). We set of three filters called accuracy filters that are: α >= 0,1, α >= 0,3, and α >= 0,5. The purpose of filtering is to see how many hypothesis, number of sentences, number of entailment pairs, and total accuracy if the filters is applied. For example on the DS-XX-R dataset, after being set of α >= 0,1 then the number of hypothesis drops to 134, which means there are 79 hypothesis with accuracy value is below 0,1. The values of number of sentences, number of entailment pairs, and total accuracy following the value of number of hypothesis.

The experimental results obtained depend on the dataset. Different dataset gives different result. Many factors may affect the results i.e data availability on Web, text extraction failures, EDA (Entailment Decision Algorithm) failures, and differences with manual checking results. Nevertheless, these failure factors can be research opportunities related to the development of RTE dataset. In addition, we also believe that our approach was feasible to generate entailment pairs that can serve as test data for textual entailment-based systems.

TABLE V. NUMBER OF ENTAILMENT PAIRS ON DS-XX-R

A	B	C	D
213	10.670	1.577	14,77%
134 (α>=0,1)	5.047	1.338	26,51%
70 (α>=0,3)	1.478	769	52,02%
35 (α>=0,5)	689	481	69,81%

[A.] Number of hypothesis; [B.] Number of sentences; [C.] Number of entailment pairs; [D.] Total of accuracy ∑α;

TABLE VI. NUMBER OF ENTAILMENT PAIRS ON DS-XX-W

A	B	C	D
74	3.806	299	6,40%
25 (α>=0,1)	682	149	21,84%
8 (α>=0,3)	112	61	54,46%
6 (α>=0,5)	65	43	66,15%

[A.] Number of hypothesis; [B.] Number of sentences; [C.] Number of entailment pairs; [D.] Total of accuracy ∑α;

VI. CONCLUSION AND FUTURE WORK

This paper describes the development of INARTE, an Indonesia dataset for research purposes in the field of Recognition Textual Entailment (RTE). Development of RTE datasets other than English has been widely practiced but our knowledge for Indonesian has never been done, therefore we trying to focus on this topic. In this paper there are three stages in making INARTE i.e (1) Collecting question-answer pairs; (2) Retrieving text from Web; and (3) Making text-hypothesis pairs. Furthermore, the paired of text-hypothesis will be evaluated whether it has entailment value or not using a subset technique. The experimental results show that our approach is able to produce 1,577 text-hypothesis pairs. Though the results have not been compared with manual annotations, we believe that the half of number will be same.

The future work of this research will continues to the topic of QAS validation using RTE. We used DS-XX-W dataset aims to see whether the non-fact sentence also gives entailment pairs. If so, then the entailment pairs may affect the answer validation process. Ideally, the test results of DS-XX-W is not expected provide entailment pairs but the reality is not. Therefore this issue will be challenge in the next research.

REFERENCES

[1] Dagan, I., Roth, D., Sammons, M., & Zanzotto, F. M, "Recognizing textual entailment: Models and applications", Synthesis Lectures on Human Language Technologies, 6(4), pp. 1-220, 2013.

[2] Bentivogli, L., Dagan, I., & Magnini, B, "The Recognizing Textual Entailment Challenges: Datasets and Methodologies", In Handbook of Linguistic Annotation, Springer, Dordrecht, pp. 1119-1147, 2017.

[3] Penas, A., Rodrigo, A., & Verdejo, F, "Sparte, a test suite for recognising textual entailment in spanish", In International Conference on Intelligent Text Processing and Computational Linguistics, Springer, Berlin, Heidelberg, pp. 275-286, 2006.

[4] Dagan, I., & Glickman, O., "Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability", In PASCAL workshop on

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on January 25,2023 at 13:55:10 UTC from IEEE Xplore. Restrictions apply.

Learning Methods for Text Understanding and Mining, Grenoble, France, (pp. 26-29), 2004.

[5] Dagan, I., Glickman, O., & Magnini, B, "The PASCAL recognising textual entailment challenge", In Machine learning challenges: evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, Springer, Berlin, Heidelberg, pp. 177-190, 2006.

[6] Alabbas, M., "A Dataset for Arabic Textual Entailment", In Proceedings of the Student Research Workshop associated with RANLP 2013, pp. 7-13, 2013.

[7] Zeller, B. D., & Pado, S, "A search task dataset for German textual entailment", In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers, pp. 288-299, 2013.

[8] Neverilova, Z., "Paraphrase and textual entailment generation", In International Conference on Text, Speech, and Dialogue, Springer, Cham, pp. 293-300, 2014.

[9] Kiss, T., & Strunk, J., "Unsupervised multilingual sentence boundary detection", Computational Linguistics, 32(4), 485-525, 2006.

Authorized licensed use limited to: Institut Teknologi Bandung. Downloaded on January 25,2023 at 13:55:10 UTC from IEEE Xplore. Restrictions apply.