Question analysis for Indonesian comparative question

To cite this article: A Saelan et al 2017 J. Phys.: Conf. Ser. 801 012077

View the article online for updates and enhancements.
Question analysis for Indonesian comparative question

A Saelan\textsuperscript{1}, A Purwarianti\textsuperscript{2}, D H Widyantoro\textsuperscript{3}

School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jl. Ganesa No. 10, Bandung, Indonesia

\textsuperscript{1}athiasaelan@yahoo.co.id, \textsuperscript{2}ayu@stei.itb.ac.id, \textsuperscript{3}dwi@stei.itb.ac.id

Abstract. Information seeking is one of human needs today. Comparing things using search engine surely take more times than search only one thing. In this paper, we analyzed comparative questions for comparative question answering system. Comparative question is a question that comparing two or more entities. We grouped comparative questions into 5 types: selection between mentioned entities, selection between unmentioned entities, selection between any entity, comparison, and yes or no question. Then we extracted 4 types of information from comparative questions: entity, aspect, comparison, and constraint. We built classifiers for classification task and information extraction task. Features used for classification task are bag of words, whether for information extraction, we used lexical, 2 previous and following words lexical, and previous label as features. We tried 2 scenarios: classification first and extraction first. For classification first, we used classification result as a feature for extraction. Otherwise, for extraction first, we used extraction result as features for classification. We found that the result would be better if we do extraction first before classification. For the extraction task, classification using SMO gave the best result (88.78\%), while for classification, it is better to use naïve bayes (82.35\%).

1. Introduction

Human demand on information seeking is increasing progressively. More people used search engine to seek for information, whether for educational purpose, or even to obtain information about goods to be bought. One of the active researches on this field is question answering. With question answering system, people can get the answer of a question directly, rather than read all of the passage retrieved by search engine.

One purpose of information seeking, especially when looking something to be bought, is to compare several objects. If we used search engine, we have to search each object manually and read some information that we want to compare. This is clearly more inconvenient than searching for only one thing.

In this research, we will develop a question answering system for answering comparative questions. This type of question answering certainly required a different question analysis compared to common factoid question answering system. This paper will explain question analysis part of comparative question answering system. This will be done by classification and information extraction technique. The answer finder component will be done in the following step of the research

Research of question answering for Bahasa Indonesia have been done before, including for factoid questions \cite{1}\cite{2}\cite{3}, non-factoid questions \cite{2}\cite{4}, and list factoid questions \cite{5}. Some research focused on the question analysis including \cite{6}, which is an Indonesian question analysis for cross language question answering. Meanwhile, research in \cite{7} focused on question analysis for complex question,
which is a question that can be decomposed into several simple questions. However, as far as we know, there is no research for comparative question answering yet.

Next part of this paper will continue as follows. Section 2 explains some related researches on information extraction from comparative sentences. Section 3 contains question analysis for Indonesian comparative question. Section 4 explains experiment and the result, while the last will conclude the result obtained from this research.

2. Related work
Several researches had studied information extraction techniques for analyzing comparative sentences. Some of the research analyzed comparative sentences for sentiment analysis [8][9][10]. Information extraction in [8] was used for extracting 3 comparative elements, that is subject entity, comparative predicate, and object entity. In [9], extracted information including 2 compared entities, relation word, and feature that is compared. Research in [10] studied how to determine preferred entities among the 2 compared entities that had been extracted in [9].

In other work, [11] also focused on comparative question. On that research, comparative questions, that was collected online, was used to get entities that is compared, or in other word comparable entities. Those comparable entities could be used for several purposes, such as recommendation systems.

3. Comparative question analysis
Comparative question is a question that comparing two or more entities, for example “bagusan mana Galaxy Grand 2 sama lenovo p70?” (English: “which one is better, Galaxy Grand 2 or lenovo p70?”). Unlike [11], we tried not to limit comparative question as questions that explicitly mentioned the compared entities. For example, “saingan berat alcatel flash 2 ini apa ya?” (English: “what is the competitor of alcatel flash 2?”) mentioned just one entity. However that sentence was a comparative question because it compared the entity to other entities. Furthermore, the sentence “Uang 1.6 jt bisa dpt hp sprti apa ya?” (English: “what kind of hand phone can be bought with 1.6 million rupiahs?”) was not mentioning any entities, but it was a comparative question because it was intended to compare all hand phones that could be bought with a certain amount of money.

For analyzing comparative questions, first we classified questions based on their type and source of the answer. Based on analysis of collected question examples, we classified comparative questions into 5 categories, which will be discussed in section 3.2. Before that, we will explain types of information that need to be extracted in section 3.1.

3.1. Information need to be extracted
There are 4 types of information to be extracted from a comparative question, including entity, aspect, comparison, and constraint. Those information were considered to be enough to represent a comparative question. For example, the sentence “xiaomi yg kualitasnya setara sama redmi note 3 yg support micro sd tipe apa ya?” (English: “which xiaomi smartphone that have similar quality with redmi note 3 and have micro sd support?”). Extracted information from that sentence are:

- Entity, i.e. the compared object. In that sentence, “redmi note 3”.
- Aspect, i.e. the thing that is compared. In that sentence, “quality”.
- Comparison. In that sentence, “similar”.
- Constraint. In that sentence two constraints are mentioned, “xiaomi” and “micro sd support”. The constraint “xiaomi” might be easily confused for an entity, but “xiaomi” here is considered as constraint because the expected answer for that question is limited to the “xiaomi” branded handphone.

Based on those information, we could understand what was asked, that is to search for an object with “similar” ”quality” with “redmi note 3”, “xiaomi” brand, and have “micro sd support”. Even
though so, this four type of information is not always exist in a question. For example, the sentence “Uang 1.6 jt bisa dpt hp sperti apa?” (English: “what kind of handphone can be bought with 1.6 million rupiahs?”) only contain the constraint. This depends on the type of the question, which will be explained in the next section.

3.2. Question types

Based on type of answer, questions can be grouped into 3, which is question answered with entity (selection), comparison, and yes/no question. Question answered with entity can be further divide into 3 types based on source of the answer which is from entities mentioned in the question, from entities not mentioned in the question, and from any entity, whether it was mentioned or not. So we have 5 question types:

1. Selection between mentioned entities (mentioned)
2. Selection between unmentioned entities (unmentioned)
3. Selection between any entities (any)
4. Comparison
5. Yes or no

In the first type, selection between mentioned entities, the expected answer is one of the mentioned entities, for example “bagusan mana Galaxy Grand 2 sama lenovo p70?” (English: “which is better, Galaxy Grand 2 or lenovo p70?”). This type of question definitely contains entity. Furthermore, the question could contain aspect, for example in the question “kalau Kamarenya bagusan mana dengan Sony 13 mp dengan Asus 13 mp” (English: “how about the camera? Which one is better compared to Sony 13mp and Asus 13mp”), the aspect is camera. If there was no information about aspect, we could assume that the comparison made is general. While for the comparison, majority of the question asked about better entity among all mentioned entities.

For the unmentioned type, the question expects an answer from unmentioned entity. There are two majority questions in this type. The first asks about similar entity with a certain constraint, for example “xiaomi yg kualitasnya setara sama redmi note 3 yg support micro sd tipe apa ya?” (English: “which xiaomi smartphone that have similar quality with redmi note 3 and have micro sd support?”). The other one asked for entities that had some better quality to the mentioned, for example “Stuart Hughes iPhone 5 Black Diamond masih biasa aja, ada yg lbih mahal ga?” (English: “Stuart Hughes iPhone 5 Black Diamond is just so-so, isn’t there anything more expensive?”). This type of question can contain all kinds of information, including entity, comparison, aspect, and constraint, like the first example. But the mandatory thing here is entity and comparison, for example the entity in the second example is “Hughes iPhone 5 Black Diamond” and the comparison is “more expensive”.

Answer expected from the third question type is any entity, mentioned or not, as long as it meets the constraints and comparisons. However, most of this type of question was not explicitly mention the comparison, but it could be assumed that people will look for for the best thing with certain constraints. For example, in the question “Uang 1.6 jt bisa dpt hp sperti apa?” (English: “what kind of handphone can be bought with 1.6 million rupiahs?”), the asker was certainly looking for the best handphone that can be bought with a certain amount of money and was not looking for the worse even though the price is far below that mentioned amount.

Questions in comparison type requires an answer in the form of comparison between two or more entities. For example, “pngen tw bedanya redmi note 2 prime sama yg redmi note 2 biasa apa ya?” (English: “what is the difference between redmi note 2 prime and the ordinary redmi note 2?”). Unlike selection between mentioned entities, this type of question doesn’t require any selected entity in the answer, just the comparison. Information contained in this type of question usually is just some compared entities. However, it is possible that there is only 1 entity, means to compare it to another entities generally. Furthermore, the question can also contain aspect, which means comparing only the selected aspect.
The last one is yes or no questions. Answer expected from this type of question is whether yes or no. For example, “semuanya sama ya dengan xiaomi redmi 2? hanya beda di ram dan internal aja kan?” (English: “it is the same as xiaomi redmi 2, right? The only difference is the ram and the internal, right?”). This type of question was more varied. Some asked the truth of a comparative sentence between some entities and with a comparative word, which would be similar to “selection between mentioned entities”. While the other question asked whether there was exists some entity with specific constraints, and will easily confused with comparison class.

4. Experiment
This experiment was aimed to find out which scenario and algorithm is suitable for comparative question analysis. In this experiment, we used Weka as machine learning tools. We compared 2 scenario. First one, we did classification first and then extraction. And the last one we did extraction first, and then classification.

4.1. Experiment Data
Data was collected from several Indonesian gadget review sites. We have collected comments from 6 sites, that is www.detekno.com, www.ulasgadget.com, hargahpxiaomi.com, ulashape.com, www.begawei.com, and hpsaja.com. Gadget review sites were chosen because we can find comparative questions quite easily from its comments. Besides, we add some questions from a general forum, kaskus.co.id, because we found that gadget review site does not have many question classified as “selection from other entities”.

Table 1. Statistics of classification data

| Question Class                                      | Data Size | Training | Testing |
|----------------------------------------------------|-----------|----------|---------|
| Selection between mentioned entities               | 106       | 85       | 21      |
| Selection between unmentioned entities             | 11        | 9        | 2       |
| Selection between any entity                       | 15        | 12       | 3       |
| Comparison                                         | 14        | 11       | 3       |
| Yes-no                                              | 26        | 21       | 5       |
| **Total**                                          | **172**   | **138**  | **34**  |

Table 2. Statistics of extraction data

| Token Class          | Data Size | Training | Testing |
|----------------------|-----------|----------|---------|
| other                | 1145      | 912      | 233     |
| entity-b             | 304       | 240      | 64      |
| entity-i             | 685       | 540      | 145     |
| aspect-b             | 48        | 36       | 12      |
| aspect-i             | 16        | 16       | 0       |
| comparison-b         | 140       | 111      | 29      |
| comparison-i         | 29        | 25       | 4       |
| constraint-b         | 74        | 60       | 14      |
| constraint-i         | 116       | 100      | 16      |
| **Total**            | **2557**  | **2040** | **517** |
Before doing the experiment, the data was preprocessed manually. We selected only comparative questions from the data. Some data that have ellipsis needs to be completed by adding entity in appropriate place according to the review title. Finally, we collected 172 comparative questions with 2557 token. Table 1 shows data for question classification, and table 2 shows data for extraction.

### 4.2. Experiment Result

Features that was used for classification task is bag of words. For extraction, we used 6 features, that is lexical, 2 previous words lexical, 2 following words lexical, and previous label. For the first experiment, that is classification first, extraction feature was added by classification result, which is question class. While the second experiment, extraction result was added as classification feature, which is number of entities, aspects, comparisons, and constraints in the sentence.

We compared 4 machine learning algorithms for each experiment, including naï ve bayes (NB), SMO, j48, and random forest (RF) that is provided by Weka. The experiment result is shown in table 3.

| Scenario        | Task   | Accuracy  |
|-----------------|--------|-----------|
|                 |        | NB | SMO | J48 | RF |
| Classification first | classification | 73.53% | 73.53% | 70.59% | 67.65% |
|                 | extraction | 84.33% | 89.75% | 84.72% | 77.56% |
| Extraction first | classification | 82.35% | 76.47% | 79.41% | 76.47% |
|                 | extraction | 84.53% | 88.78% | 84.72% | 83.37% |

From the experiment, we can see that for extraction task, there was no significant difference between the two scenarios, except for random forest classification. But for classification, the second experiment gave better result. So we could say that the extraction first scenario will gave better result generally.

For classification task, naï ve bayes gave better result. Most of the prediction error occurred in yes/no class that is classified as mentioned class. This is consistent with the analysis in section 3.2.5, a question that asked the truth of a comparative sentence between some entities would be similar to “selection between mentioned entities”. This encouraged us to do some further analysis for this type of question in the future research. In the best case of the extraction task, the one that use naï ve bayes classifier, we found out that from 5 yes/no question, 3 question was misclassified as mentioned. SMO was better for classifying yes/no questions, which was only 2 questions that misclassified. But SMO was not really good compared to naï ve bayes, especially for unmentioned and comparison class.

For extraction task, we found that SMO gave better result. The worst error occurred in aspect-b class and constraint-b class. For aspect-b class, only 1 from 12 training data was classified correctly. While for constraint-b, only 5 from 14 that were classified correctly. Most of the misclassified data were classified into other class. We guessed that aspect and constraint did not give good result because of inconsistent annotation.

### 5. Conclusion

We have analyzed some Indonesian comparative question and tried to make automatization of question analysis for comparative question. We grouped comparative questions into 5 types, which is selection between mentioned entities, selection between unmentioned entities, selection between any entity, comparison, and yes or no question. Then we extracted 4 types of information from comparative questions, which is entity, aspect, comparison, and constraint. Experiment showed that it is better to do the extraction first, and then the classification task used the extraction result as features. For extraction task, classification using SMO gave the best result (88.78%), while for classification, it
is better to use naïve bayes (82.35%). For future research, we can further analyzing about yes/no questions. Besides, we will continue this research by developing answer finder for comparative questions.

Acknowledgement
This research is partially supported by "Penelitian Pendidikan Magister menuju Doktor untuk Sarjana Unggul (Kemenristekdikti)" (Master's Program toward Doctoral Degree for Excellent Graduate) within research "Pengembangan Chatbot Berbahasa Indonesia" (Indonesian Chatbot Development).

References
[1] Purwarianti A, Tsuchiya M and Nakagawa S 2007 A Machine Learning Approach for Indonesian Question Answering System Proc. Artificial Intelligence and Applications (12-14 February 2007, Innsbruck, Austria) (ACTA Press) pp 573-578
[2] Zulen A A and Purwarianti A 2011 Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question Proc. PAACLIC 25 (16-18 December 2011, Singapore) (Waseda University) pp 622-631
[3] Zulen A A and Purwarianti A 2013 Using Phrase-Based Approach in Machine Learning Based Factoid Indonesian Question Answering Proc. CISAK (7 July 2013, Daejon, South Korea)
[4] Purwarianti A and Yusliani N 2010 Study of Non Factoid Question Answering for Bahasa Indonesia Proc. ICACSIS 2010 (20-23 November 2010, Bali, Indonesia)
[5] Prabawa A E and Purwarianti A 2011 Indonesian List Question Answering System. Proc. KNIF 2011 (23 November 2011, Bandung, Indonesia)
[6] Faruqi M I and Purwarianti A 2011 An Indonesian Question Analyzer to Enhance the Performance of Indonesian-English CLQA Proc. ICEEI 2011 (17-19 July 2011, Bandung, Indonesia) (IEEE)
[7] Basuki S and Purwarianti A 2016 Statistical-based approach for Indonesian complex factoid question decomposition JIEEE 8(2) 356
[8] Yang S and Ko Y 2011 Extracting Comparative Entities and Predicates from Texts Using Comparative Type Classification Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (19-24 June 2011, Portland, Oregon) vol 1 (ACL) pp 1636-1644
[9] Jindal N and Liu B 2006 Mining Comparative Sentences and Relations Proc. AAAI 2006 (16-20 July 2006, Boston, Massachusetts) 22 (Menlo Park, California: AAAI Press) pp 1331-1336
[10] Ganapathibhotla M and Liu B 2008 Mining Opinions in Comparative Sentences Proc. 22nd International Conference on Computational Linguistics (18-22 August 2008, Manchester, UK) vol 1 (ACL) pp 241-248
[11] Li S, Lin C Y, Song Y I and Li Z 2013 Comparable Entity Mining from Comparative Questions IEEE Transactions On Knowledge And Data Engineering 25(7) pp 1498-1509