Indonesian-English Cross Language Question Answering

Ayu Purwarianti, Masatoshi Tsuchiya, Seiichi Nakagawa
Department of Information and Computer Sciences, Toyohashi University of Technology,
ayu@slp.ics.tut.ac.jp, tsuchiya@imc.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp

Abstract

Our Indonesian-English Cross Language Question Answering (CLQA) is divided into 4 components: question analyzer, keyword translator, passage retriever and answer finder component. The Indonesian question is inputted into a question analyzer which yields Indonesian keyword list, Indonesian question focus and question class. We defined the question class by using an SVM machine implemented in Weka[15]. Because Indonesian is a poor data resource language, we use a bigram frequency feature as an addition feature for the question classification. The Indonesian keywords are translated into English using an Indonesian-English bilingual dictionary. The English translations are composed into a boolean query to retrieve relevant passages. We select the passages within 3 highest IDF scores. In the answer finder, the answer is located by using an SVM method for text chunking implemented in Yamcha[4]. Different with other Indonesian-English CLQA[1,14], we do not tag the name entities in the target documents, instead we only do the POS tagging by using TreeTagger[12] for the target documents. Based on our experiment in Indonesian QA, we choose to use question class, question features and document features for the machine learning based answer finder. We also complement the WordNet distance feature for the document features. By using 284 questions as the test data, we achieved about 31.69% accuracy on top 5 answers which is better than other Indonesian-English CLQAs.

1. Introduction

Cross Language Question Answering (CLQA) has been an interesting area as the extended work of a Question Answering (QA). CLEF (Cross Language Evaluation Forum) started the CLQA task since 2003 by providing English documents for Italian, Spanish, Dutch, French and German queries[6]. Indonesian-English CLQA has also become one of the tasks in CLEF since 2005[13]. NTCIR (NII Test Collection for IR Systems) has also provided the CLQA task since 2005[10] for Chinese, Japanese and English languages. In CLQA, given a question in a source language, the answer is searched in documents of target language. The accuracy of the system is measured by its retrieved correct answer.

A CLQA has a slightly different problem compared with a monolingual QA system. Sasaki, et al.[10] mentioned that the translation phase caused a CLQA more difficult than a QA on following reasons:

1. Translated questions are represented with different expressions than those used in news articles in which answers appear.
2. Since key words for retrieving documents are translated from an original question, document retrieval in CLQA becomes much more difficult than that in monolingual QA.

The data source alternatives in translating the questions or the documents are bilingual dictionary, machine translation and parallel corpus. The results on NTCIR 2005[10] showed that systems using bilingual dictionary achieved better accuracy than the machine translation approach. We assumed that this is caused by the fact that the result of a machine translation could be inadequate for covering the correct translation while the bilingual dictionary translation gives more than one translation candidate which could cover the correct translation in the retrieved passage. Bilingual dictionaries are also more available than the machine translation and the parallel corpus. We believe that this is one of the reasons of the low accuracy result of Indonesian-English CLQAs [1,14] which use a machine translation software to translate Indonesian query into English. Even though, we use a small size Indonesian-English dictionary[2] (29,054 word entries), our CLQA system gives better result than other Indonesian-English CLQA systems. We define that our Indonesian-English dictionary is a small size dictionary by comparing it to the three dictionaries (EDICT 110,428 entries, ENAMDICT 483,691 entries and in-house translation dictionary 660,778 entries) used by Isozaki, et al.[5]. The experiment result of Isozaki, et al.[5] achieved the highest performance of 31.5% accuracy (top 1 answer) for exact answer in the Japanese-English task of NTCIR 2005.

Another point that we want to emphasize is the answer finder module. Mostly CLQA systems extract the answer by matching the named entity of the answer candidate and the question class which is predicted from the question[1,5,11,14]. Here, we do not do the named entity tagging for the documents, instead we try to locate the answer by a text chunking process. Our approach is almost similar to Sasaki[9]. The differences between our approach and Sasaki[9] are explained in the Section 4.4 (Answer Finder).

The rest of the paper is organized as follows. Section 2 presents the related works. Section 3 describes the language resource used in our Indonesian-English CLQA. Section 4 discusses each component in our Indonesian-English CLQA. Section 5 shows our experimental result using in-house test data and the question set from NTCIR 2005 CLQA task. Section 6 describes our research conclusion and our next research plan.

2. Related Works

Here, we would like to show two other works in Indonesian-English CLQA system. Both are conducted for the QA@CLEF task. First, Adriani&Rinawati[1] in CLEF 2005 used a ruled based Indonesian question categorizer, a commercial Indonesian-English machine translation software (Transtool), Lemur information retrieval system and calculate the passage (one passage is two sentences) score based on the similarity of the named entity of English documents which have been tagged by Monty Tagger. They reported that the Transtool only failed to translate 8 Indonesian words. But the answer finder results showed low accuracy score (right: 2, inexact: 36, unsupported: 0, wrong: 162, of 200 questions as the test data).

The second system is for the CLEF 2006. Wijono, et al.[14] used a ruled based question classifier, Kataku machine translation tool, Lemur information system and Gate named entity tagger. The answer is the one with appropriate tag and has the smallest distance with the query word found in the passage. The answer finder gave better result than the previous year (right: 14, inexact: 4, unsupported: 13, wrong: 159, of 200 questions as the test data).

Another related work is the research by Sasaki[9] as one of the participant in the CLQA1 of NTCIR 5 for the J/E and E/J tasks. Sasaki used an extended QBTE (Question Biased Term Extraction) approach which eliminates the question classification process and the named entity tagger. QBTE is a kind of statistical question answering approaches. Sasaki[9] employed the machine learning technique, Maximum Entropy Models (MEMs) to extract answers from combined features of question features and document features. The training data was 300 question answer pairs and the test data was 200 questions. For the translation phase, they used Japanese-English (24,805 word entries), English-Japanese (17,571 word entries) word dictionaries and a Japanese-English POS dictionary. In the Japanese-English CLQA task, the system resulted only 2 correct answers for 200 questions test data.

3. Language Resources

Although the Indonesian-English CLQA has been one of the task in the CLEF 2005 and 2006, we decided to build our own data of Indonesian-English CLQA related with our next research on Indonesian-Japanese CLQA. By building our own Indonesian-English CLQA, we also could have data for the training phase. We asked 16 Indonesian college students to read some English articles (taken from Daily Yomiuri Shimbun year 2000) and made Indonesian questions based on those articles. After deleting the exact questions, we got 2553 questions for training data and 284 questions for test data in 6 question classes (person, organization, location, quantity, date, name). The detail collected questions number for each question class is shown in Table 1. Some question examples are shown in Table 2.

Table 1. Number of Collected Questions

Question Class	Collected Question Number
Person	447
Organization	475
Location	476
Quantity	498
Date	459
Name	482

Table 2. Examples of Inputted Questions

Question Class and Question Example
Person: Siapakah ketua Komite Penyalahgunaan Zat di Akademi Pediatri Amerika? (Who is the head of the Committee on Substance Abuse at the American Academy of Pediatrics?)
Organization: Apa nama institut penelitian yang meneliti Aqua-Explorer? (What is the name of research institute that does the experiment of Aqua-Explorer)
Location: Di provinsi manakah, Tokaimura terletak? (In which prefecture is Tokaimura?)
Quantity: Ada berapa lonceng yang terdapat di Kodo Hall? (How many bells are there in Kodo Hall?)
Name: Apa nama kartu identifikasi bagi para sopir taksi di Jepang? (What is the name of identification card for taxi driver in Japan?)

4. Method in Indonesian-English CLQA

Similar with common approach, we divide our CLQA system into four components such as question analyzer, keyword translation, passage retriever and answer finder. Following subsections will discuss each component in detail.

4.1 Question Analyzer

Our Indonesian question analyzer consists of a question shallow parser and a question classifier. We built a rule based question shallow parser to extract:

- interrogative word (such as who, when, where, etc),
- question keywords (by eliminating some stop words from the question),
- question focus (with its position against the interrogative word),
- one clue word in the question,
- and a phrase like information for the question focus or the clue word (if the question focus does not exist).

Then we calculated the bi-gram frequency scores (see [8] for detail calculation) between the question focus (or the question clue word) with some defined words (for each question class). The shallow parser results and the bi-gram frequency scores are used as the features for the SVM based question classification task. A question example along with its features for question classification is shown in Figure 1.

As has been proven in our Indonesian QA[8], using a question class in the answer finder gave higher performance than without using the question class (only depends on the question shallow parser result, mainly on the question focus). Being compared to a monolingual system, using a question class in the cross lingual QA system gives more benefits. First benefit is in the

problem if there are more than one translation for the question focus where these translations have different semantic. For example, in question "Posisi apakah yang dijabat George W. Bush sebelum menjadi presiden Amerika?" (What was George W. Bush's position before he became president of United States?), the question focus is "posisi" (position) which can be assumed as a place (*location*) or an occupation (*name*). By classifying the question into "name", then the answer extractor will automatically avoid the "location" answer. Second benefit is in the problem of an out of vocabulary question focus. By providing the question class, even though the question focus can not be translated but the answer can still be predicted using the question class.

Question: Apa nama kartu identifikasi bagi para sopir taksi di Jepang? (What is the name of identification card for taxi driver in Japan?) → question class: name

Features for question classification:

- Interrogative word: apa (what)
- Question keyword: nama (name), kartu (card), identifikasi (identification), sopir (driver), taksi (taxi), Jepang (Japan)
- Question focus: kartu (card)
- Question focus position: post (after the interrogative word)
- Phrase-like information: NP
- Existence of question focus: yes
- Bi-gram frequency score (frequency, number of words):
  - date(0,0), loc(0.0010,2), name(0.0268, 10),
  - organization(0,0), person(0.0020, 1), quantity(0,0)

Result of question classification: date: 0.00, loc: 0.20, name: 0.33, organization: 0.26, person: 0.13, quantity: 0.07 → name

Figure 1. Feature Example for the Question Classification

4.2 Keyword Translation

Based on our observation of the collected Indonesian questions, we assume that there are three types of words used in the Indonesian question sentence:

1. Native Indonesian words, such as "siapakah" (who), "bandara" (airport), "bekerja" (work), etc
2. English words, such as "barrel", "cherry", etc.
3. Transformed English words, such as "presiden" (president), "agensi" (agency), "prefektur" (prefecture), etc.

For the native Indonesian words, we use the Indonesian-English bilingual dictionary[2] (29,047 entries). For the English words, we search whether the word exists in the corpus or not. For the transformed English words, some translation candidates are defined automatically using some transformation rule such as "k" into "c", or "si" into "cy", etc. The words that exist in the English corpus are assumed as the translations. By using this schema, among 3706 unique key words in our 2837 questions, we got 153 OOV words.

4.3 Passage Retriever

The English translations are then combined into a Boolean query which use "or", "and" and "or2" operators. "or2" operator is used for synonyms (such as in [7]). By joining all the translations into a Boolean query, we do not filter the keywords into only one translation candidate such as done in machine translation method.

We retrieve the relevant passages in two steps: document retrieval and passage retrieval. For the document retrieval, we select documents with IDF score higher than the highest IDF score divide by 2. And for the passage retrieval, we select passages in the retrieved documents within the three highest IDF scores.

4.4 Answer Finder

As has been mentioned in the introduction section, we do not do the named entity tagger in the answer finder phase, instead we treat the answer finder as a text chunking process. Each word in the corpus will be given a status as a "B" or "I" or "O" based on some features of the document word and also based on some features of the question. We use an available text chunking software Yamcha that works using an SVM algorithm.

In the maximum entropy based QBTE approach, Sasaki[9] used some POS information for a word, for example the word "Tokyo" is analyzed as POS1=*noun*, POS2=*propernoun*, POS3=*location*, and POS4=*general*. The POS information of each word in the question is matched with the POS information of each word in the corpus by using a true/false score in the feature of the machine learning. In our Indonesian-English CLQA, we do not use such information. The POS information in our Indonesian-English CLQA is similar to the POS1 mentioned in Sasaki[9]. Even though our Indonesian-English dictionary has bigger size than the Japanese-English dictionary (24,805 entries) used by Sasaki, our Indonesian-English dictionary does not posses the POS2, POS3 and POS4 such as in Sasaki.

Another difference is that we use question class as one of question features with reasons such as mentioned in Section 4.1. We also use the result of our question shallow parser along with the bi-gram frequency score.

For the document features, each word is morphologically analyzed into its root word using TreeTagger[12]. The root word, its orthographic information and its POS (noun, verb, etc) information are used as the question features. Different with our Indonesian QA[8], we do not calculate bi-gram frequency score for the document word, instead we calculate its WordNet distance with 25 synsets such as listed in the noun lexicographer files of WordNet. Each document word is also complemented by its similarity scores with the question focus, question clue word and question keywords. If a question keyword consists of two word such as "territorial water" translation for "perairan", then for a document word matches with one word inside it, the score is divided by the number of words in that question keyword. For example, for document word "territorial", the similarity score against "territorial water" is 0.5. An example of the features used for the Yamcha as a text chunking software is shown in Figure 2.

Question: equal with Figure 1, answer: Shigematsu Kan
Question features: equal with Figure 1 + answer type: name
Document word features:
- lexical term: taxi - WordNet distance: 1 for artifact
- POS: NN - similarity score: 1
The retrieved passage:
.. the identification card of the taxi driver read " Shigematsu Kan " as the kanji character for his family name …
Predicted answer: Shigematsu Kan

Figure 2. Example of Features for the Text Chunking

5. Experiments

5.1 Question Classification

In the question classification experiment, we applied an SVM algorithm in WEKA software [15] with linear kernel and the "string to word vector" function to process the string value. We used 10-fold cross validation for the accuracy calculation. The accuracy result is 95.84%. The detail performance for each question class is in Table 3.

The lowest performance is for the "organization" class. For example, question "Siapa yang mengatakan bahwa 10% warga negara Jepang telah mendaftarkan diri untuk mengikuti Pemilu pada tahun 2000?" (who says that 10% of Japan citizen have applied for the national election in year 2000?) got a "person" as the classification result. Even for a human, it is quite difficult to define the question class of the above example without knowing the correct answer.

Table 3. Confusion Matrix for Question Classification

in \ out	person	org	loc	quan	date	name
person	439	8	0	0	0	0
org	29	401	7	0	0	38
loc	1	13	458	0	0	11
quan	0	0	0	497	1	0
date	0	0	0	0	459	0
name	1	13	3	0	0	465

5.2 Passage Retriever

For the passage retriever, we used two evaluation measures: precision and recall. Precision shows the average ratio of relevant documents. Relevant document is a document that contains a correct answer without considering supporting evidence. Recall shows number of questions that might have correct answer in the retrieved passages. Our passage retrieval achieves precision of 0.124 (1012 passages among 8249 retrieved passages) and recall of 0.708 (201 answerable questions among the 284 questions). For the English target corpus, we use Daily Yomiuri Shimbun year 2000 and 2001.

5.3 Question Answering Accuracy

To measure our CLQA performance, we use the Top1, Top5 and MRR scores for the exact answers, such as shown in Table 4.

Table 4. Performance of Indonesian-English CLQA

Top1	Top5	MRR	InExact
64 (22.54%)	90 (31.69%)	67.02	15 (5.28%)

6. Conclusions

Our experiment shows that for a poor resource language such as Indonesian, it is still possible to be able to build a cross language question answering with a promising result.

We found some weaknesses in our passage retrieval modules and the similarity score features in our answer finder module. For our next research plan, we will try to improve these two modules. We will also try to develop our system for an Indonesian-Japanese CLQA.

Acknowledgement

This work was partially supported by The 21st Century COE Program "Intelligent Human Sensing".

References

[1] Adriani, Mirna, and Rinawati, "University of Indonesia Participation at Query Answering-CLEF 2005", Proceedings of CLEF 2005 Workshop, 21-23 September 2005, Vienna, Austria.

[2] Agency for The Assessment and Application of Technology, KEBI (Kamus Elektronik Bahasa Indonesia), http://nlp.aia.bppt.go.id/kebi/ (last access: February 2004).

[3] Fellbaum, C., editor, "WordNet: an electronic lexical database and some of its application", MIT Press, Cambridge, MA, 1998.

[4] Kudoh, Taku Y. Matsumoto, "Use of Support Vector Learning for Chunk Identification", Proceedings of the Fourth Conference on Natural Language Learning (CoNLL-2000), Lisbon, Portugal, 2000, pp. 142-144.

[5] Isozaki, Hideki, Katsuhito Sudoh, Hajime Tsukada, "NTT's Japanese-English Cross-Language Question Answering System", Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan, pp. 186-193.

[6] Magnini, Bernardo, et.al., "The Multiple Language Question Answering Track at CLEF 2003", Proceedings of CLEF 2003 Workshop, 21-22 August 2003, Norway.

[7] Pirkola, A., "The Effects of Query Structure and Dictionary Setups in Dictionary-based Cross Language Information Retrieval", Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 55-63, 1998.

[8] Purwarianti, Ayu, Masatoshi Tsuchiya, Seiichi Nakagawa, "A Machine Learning Approach for Indonesian Question Answering System", IASTED AIA (Artificial Intelligence and Application) 2007, Innsbruck, Austria, February 2007.

[9] Sasaki, Yutaka, "Baseline Systems for NTCIR-5 CLQA1: An Experimentally Extended QBTE Approach", Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan, pp. 230-235.

[10] Sasaki, Yutaka, Hsin-Hsi Chen, Kuang-hua Chen, Chuan-Jie Lin, "Overview of the NTCIR-5 Cross-Lingual Question Answering Task (CLQA1)", Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan, pp. 175-185.

[11] Sekine, Satoshi and Ralph Grishman, "Hindi-English Cross Lingual Question-Answering System", ACM Transactions on Asian Language Information Processing, Vol. 2, No. 3, September 2003, pp. 181-192.

[12] Universitat Stuttgart, Institute for Natural Language Processing, "TreeTagger Project", http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ (last access: December 2006).

[13] Vallin, Alessandro, et al., "Overview of the CLEF 2005 Multilingual Question Answering Track", Proceedings of CLEF 2005 Workshop, 21-23 September 2005, Vienna, Austria.

[14] Wijono, Sri Hartati, Indra Budi, Lily Fitria, Mirna Adriani, "Finding Answers to Indonesian Questions from English Documents", Proceedings of CLEF 2006 Workshop, 20-22 September 2006, Alicante, Spain.

[15] Witten, Ian H. and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques 2nd edition", Elsevier Inc., 2005.