Finding Answers to Indonesian Questions from English Documents

Sri Hartati Wijono, Indra Budi, Lily Fitria, and Mirna Adriani

Faculty of Computer Science
University of Indonesia
Depok 16424, Indonesia
{shw50, indra, lifi50, mirna}@cs.ui.ac.id

Abstract. Our report describes the results of work in our participation in the Indonesian-English question-answering task of the 2006 Cross-Language Evaluation Forum (CLEF). In this work we translated an Indonesian query set into English using a machine translation tool available on the internet. Documents relevant to a question are first retrieved. The relevant documents are then divided into passages of 5 sentences each. The answer to the question is extracted from a passage relevant to the query. The answer is identified based on matching annotations between the query and the documents. A linguistic tool is used to annotate the words in the queries and the documents.

1 Introduction

This year we participate in the bilingual Question-Answering (QA) task, the Indonesian-English QA, of the Cross Language Evaluation Forum (CLEF) 2006. Finding the correct answer to a question in documents is a challenging task, and this is the main research topic in CLEF Question Answering task. The question and the documents must be analyzed in order to find the right answer in the documents. There are several techniques that have been used to handle the QA task, i.e., parsing and tagging the sentence in the question [6], in the documents [4], in paragraphs [8], and in passages [2, 3].

2 The Question Answering Process

The process of finding the answer to a query in documents proceeds in a number of stages. First, the original CLEF English queries were translated into Indonesian manually. Next, we classified the Indonesian questions (queries) according to the type of question. We identified the question type from the question word used in the query. The Indonesian question was then translated back into English using a machine translation tool. We used a web machine translation tool called *Toggletext*[1] to translate an Indonesian query set into English. We learned from our previous work [1] that freely available dictionaries on the Internet did not provide sufficiently good translation terms, as their vocabulary was very limited. We hoped that we could achieve better results using a machine translation approach.

The resulting English query was then used to retrieve the relevant documents from the collection by means of an information retrieval system. The contents of a number of documents at the top of the rank list were then split into passages. The passages were then tagged using a linguistic tagging (annotation) tool to identify the type of words in the passages. Finally, the passages were then scored using an algorithm, and the answer to the question is extracted from the passage with the highest score.

2.1 Categorizing the Questions

Each question category, which is identified by the question word in the question, points to the type of answer that is looked for in the documents. The Indonesian question-words used in the categorization are:

dimana, dimanakah, manakah (where)	points to <location>
apakah nama (what),	points to <location>

[1] See "http://www.toggletext.com".

siapa, siapakah (who)	points to <person>
berapa (how many)	points to <measure>
kapan (when)	points to <date>
organisasi apakah (what organization)	points to <organization>
apakah nama (which)	points to <location>
sebutkan (name)	points to < other>

By identifying the question type, we can predict the kind of answer that we need to look for in the document. The Indonesian question was tagged using a question tagger that we developed according to the question word that appears in the question. This approach is similar to those used by Clark et al. [2] and Hull [4,7,8]. However, we ignored the tagging on the question when we ran the query through the IR system to retrieve the documents.

2.2 Building Passages

The Indonesian question was translated into English using machine translation. The resulting English query was then run through an information retrieval system as a query to retrieve a list of relevant documents. We used *Lemur*[2] information retrieval system to index and retrieve the documents. The contents of the top 10 relevant documents were split into passages. Each passage contains five sentences where the last sentence is repeated in the next passage as the first sentence. The sentence in the documents was identified using a sentence parser to identify the beginning and the end of a sentence. The passages are then indexed by Lemur and the queries were run through to get the top-10 passages. These top-10 passages are being scored in order to get the answers to the queries.

2.3 Tagging the Passage

The passages were then run through an entity tagger to get the entity annotation tags. The entity annotation tagger identifies words of known entity types, and tags them with the entity type tags, such as person, location, and organization. For example, <*organization*> *UN*, the word *UN* is identified as an organization so it gets the organization tag. In this work, we used linguistic tagger tool, *Gate*[3].

Gate analyzes English words and annotates them with tags to indicate location, organization, and person, where applicable. The annotation tags are used to find the candidate answer based on the type of the question, for example, a word with location tag is a good candidate answer to a *where* question, and a word with a person tag is a good candidate answer to a *who* question.

2.4 Scoring the Passages

Passages were scored based on their probability of answering the question. The scoring rules are as follows:

1. Give 1 to a passage if its tag is not the same as the query tag and 0 if not.
2. Add 1 if a word in the passage is the same as the query.
3. Add 1 if the number of words in the passage is more than half of the number of the query words.

Once the passages obtained their scores, the top 10 scoring with the appropriate tags – e.g., if the question type is person (the question word "*who*") then the passages must contain the person tag – were then taken to the next stage.

2.5 Finding the Answer

The top 10 passages were analyzed to find the best answer. The probability of a word being the answer to the question is inversely proportional to the number of words in the passage that separate the candidate word and the word in the query. For each word that has the appropriate tag, its distance from a query word found in the passage is computed. The candidate word that has the smallest distance is the final answer to the question.

---

[2] See "http://www.lemurproject.org/".
[3] See "http://www.gate.shef.ac.uk/".

For example:

- Question: What is the **capital** of <*LOCATION*> Somalia?
- Passage:
  - Here there is no coordination. <*PERSON*> Steffan de Mistura – UNICEF representative in the Somali **capital**, <*LOCATION*> **Mogadishu**, and head of the anti-cholera team – said far more refugees are crowded together here without proper housing or sanitarian than during the <*LOCATION*> **Somalia** crisis. And many are already sick and exhausted by the long trek from <*LOCATION*> **Rwanda**.

The distance between the question word *capital* and *Mogadishu* is 1, between the question word *capital* and *Rwanda* is 38. So, *Mogadishu* becomes the final answer since its distance to the question word *capital* is the smallest one (closest).

3 Experiment

Our work focused on the bilingual task using Indonesian questions to find answers in English documents. The Indonesian questions were obtained by manually translating the English questions. The Indonesian questions were then translated back into English an online machine translation tool *Toggletext* to retrieve relevant English documents from the collection.

Using the *Gate* tagger to tag words in the passages, only 14 correct answers were found (see Table 1). There were 4 inexact (ambiguous) answers and 159 wrong answers.

Table 1. Evaluation of the QA result

Task : Bilingual QA	Evaluation
W (wrong)	159
U (unsupported)	13
X (inexact)	4
R (right)	14

Our result shows that we need to find ways to improve the effectiveness in finding correct answers, in particular, ways of reducing the number incorrect word tagging. We also learned that expanding the translated queries by adding related terms could also help, as more relevant documents can be retrieved for the QA algorithm to work.

4 Summary

Our participation in the QA task still needs further improvement. In our recent work, we managed to improve the QA result by applying query expansion and different passage scoring techniques. We hope that applying such techniques will result in a better performance next year.

References

1. Adriani, M. and van Rijsbergen, C. J. Term Similarity Based Query Expansion for Cross Language Information Retrieval. In Proceedings of Research and Advanced Technology for Digital Libraries (ECDL'99). Springer Verlag, Paris (1999) 311-322
2. Clarke, C. L. A., Cormack, G.G., Kisman, D. I. E. and Lynam, K. Question Answering by Passage Selection. In NIST Special Publication: The 9th Text retrieval Conference (2000)
3. Clarke, Charles L.A., Cormack, Gordon V. and Lynam, Thomas R. Exploiting Redundancy in Question Answering. In Proceeding of ACM SIGIR. New Orleans (2001)
4. Hull, David. Xerox TREC-8 Question Answering Track Report. In NIST Special Publication: The 8th Text Retrieval Conference (1999)
5. Li, Xiaoyan dan Croft, Bruce. Evaluating Question-Answering Techniques in Chinese. In NIST Special Publication: The 10th Text Retrieval Conference (2001)
6. Manning, C.D. and Schutze, H. Foundations of Statistical Natural Language Processing. The MIT Press, Boston (1999)

7. Moldovan, D. et.al. Lasso: A Tool for Surfing the Answer Net. In NIST Special Publication: The 8th Text Retrieval Conference (1999)
8. Pasca, Marius and Harabagiu, Sanda. High Performance Question Anwering. In Proceeding of ACM SIGIR. New Orleans (2001)