Finding Answers Using Resources in the Internet

Septian Adiwibowo and Mirna Adriani

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

Abstract. In this paper we describe our experiments in finding answers from documents based on statistical and linguistic knowledge. We collected the candidate answers from sources available on the internet, and then we used them to validate the answers found in the documents. The candidate answers from the documents were found using a statistical technique and linguistic knowledge such as named entity tags to find the type of answer that matches the question category.

Keywords: question answering, query expansion.

1 Introduction

In our participation in the Question Answering task [1, 2] of Cross Language Evaluation Forum (CLEF) 2007, i.e., for Indonesian-English, we needed to use language resources to translate Indonesian queries into English. Luckily we found a machine translation tool available on the Internet that could be used to translate Indonesian queries into English.

We also made use of the information sources available on the Internet [3] to validate answers that were found in the documents of a collection. We used statistical technique to find the answers in the documents.

2 The Process of Analyzing the Questions

A number of steps were performed to the questions that we received from CLEF. Since there were only English questions, we manually translated the 200 original English questions from CLEF into Indonesian.

The query-answering process proceeds in the following stages:

1. Question categorization
2. Passages identification/building

3. Passages scoring
4. Answers identification.

First we categorize the Indonesian question according to the type of question. We identify the question type based on the question word found in the query.

The Indonesian question is then translated into English using a machine translation tool. The resulting English query is then used to retrieve relevant documents from the collection through an information retrieval system. The contents of a number of documents at the top of the list are then split into passages. The passages are then scored using an algorithm, and the passage with the highest score is chosen to be the answer to the question.

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>
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>

By identifying the question type, we can predict the kind of answer that we need to look for in the document. The Indonesian question is 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. and Hull [2, 3].

2.2 Building Passages

Next, the Indonesian question is translated into English. The resulting English query is then run through an information retrieval system as a query to retrieve a list of relevant documents. We use *Lemur*[1] information retrieval system to index and retrieve the documents. The contents of the top 50 relevant documents are split into passages. Each passage contains 100 words. The passages are then tagged using GATE (http://www.gate.shef.ac.uk/).

[1] See http://www.lemurproject.org/.

2.3 Scoring the passages

Passages are scored based on their likeliness to answer the question. The scoring rules consider the number of words from the questions that appear on the passages. Then the distance of the answer candidates and the words appear on the query are also considered.

Once the passages obtained their scores, the top 20 passages with the highest scores and have the appropriate tags – e.g., if the question type is person (the question word "*who*") then the passages must contains the person tag – are then taken to the next stage.

2.4 Finding the answer

The top 20 passages are analyzed to find the best answer. The likeliness of a word to be 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, 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. We also validate the answer candidates to the answer that we find on available sources on the internet. We get the top 50 answers for each question from Google (http://www.google.com). We then rank the words according to their word frequencies. The word that has the highest frequency is the answer candidate to a question. We then add a weight to the final score of the answer find in the document.

3 Experiment

We participated in the bilingual task with English topics. The query translation process was performed fully automatic using a machine translation technique. The machine translation technique translates the Indonesian queries into English using Toggletext[2], a machine translation that is available on the Internet. In these experiments, we used Lemur[3] information retrieval system which is based on the language model to index and retrieve the documents.

4 Results

Our work is focused on the bilingual task using Indonesian questions to retrieve answer from an English document collection. Table 1 shows the result of our experiments.

---

[2] See http://www.toggletext.com/.
[3] See http://www.lemurproject.org/.

Table 1. The QA results.

Task : Bilingual QA	Evaluation
W (wrong)	175
U (unsupported)	1
X (inexact)	4
R (right)	20

Changes in the question types this year had an impact on the number of answers that we managed to find. The scoring and the answer patterns that we identified in the previous year's questions did not work very well for this year's questions. The percentage of correct answers that we got this year was only 10%.

5 Summary

We learned from our work that using information from sources available on the internet can help verify the answers found in documents. However, deeper linguistic knowledge needs to be considered to get an even better result.

References

1. Clarke, C. L. A., G. G. Cormack, D. I. E. Kisman and K. Lynam. Question Answering by Passage Selection: *The 9th Text retrieval Conference (TREC-9)*. 2000.
2. Hull, David. Xerox TREC-8 Question Answering Track Report: *The 8th Text Retrieval Conference (TREC-8)*. 1999.
3. Hildebrandt, W., Katz, B., & Lin, J. Answering definition questions with multiple knowledge sources. *Proceedings of the 2004 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annnual Meeting (HLT/NAACL 2004)*, 2004.