Accuracy of the AID System's Information Retrieval in Processing Huge Data Collections

Jolanta Mizera-Pietraszko

Department of Information Systems, Institute of Applied Informatics
Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, Building A1, Room 203, 50-370 Wroclaw, Poland
jolanta.mizera-pietraszko@pwr.wroc.pl, lutynia@go2.pl

Abstract

The name of the AID system stands for Answer Identifier in terms of open-domain Polish-English question answering system. Both the matching technique and indexing rely on simple question taxonomy. The concept of AID developed from Quest and Aranea which are both simple question answering software. In addition, AID employs direct MT model based on phrase-for-phrase translation having incorporated the LEC multilingual component. The system architecture heuristics include extracting passages, part-of-speech tagging, semantic relations and some parsing rules rather than exploring strategies for query reformulation which were found ineffective. Finally, the system's performance in the run submitted is evaluated on the British and American data collections.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; - Performance Evaluation H.2.3 [Database Management]: Languages - Query Languages

General Terms

Languages, Measurement, Performance, Experimentation

Keywords

Question answering, Question classes

1 Introduction

In literature question answering is found as a more complex form of natural language processing. In this respect this is a step beyond information retrieval. In addition to document retrieval such a system is expected to browse the given local collection for precise and complete information in order to form a phrase or a sentence being an unambiguous response to the question posed.
AID was developed at Wroclaw University of Technology on the base of Swiss Quest and Aranea produced at the MIT Computer Science and Artificial Intelligence Laboratory in Massachusetts, USA which was tested on TREC data collections. The system works on the Windows and the Linux platforms as well. In order to make the system platform-independent, the tools Knoppix and VPC2004 were used. Despite not residing on the HDD, the newest version of Knoppix embraces drivers for reading files and saving results while booting it from PC with Windows installed.

The approach presented in this paper is intended to be both simplistic and pragmatic. The backbone of the AID system is classification of the questions both at the translation and the analysis stages. Since this system participates in the CLEF 2006 experiment for the first time, some of its components are planned to be expanded so as to improve the performance on the larger CLEF collections. Having tested a number of information retrieval systems it came clear that producing a system prototype takes considerably less time than searching for and then studying the systems' instructions. Therefore this system has been constructed to serve the purpose of the CLEF experiment.

2 The Architecture of AID

Figure 1 presents the system architecture. As seen, the relationship between its components are organized into the whole process due to the CLEF guidelines. Although the system structure may seem a little complex, in fact it reflects the most common components of any question answering system. Modularity represents the phases of the experiment in every detail.

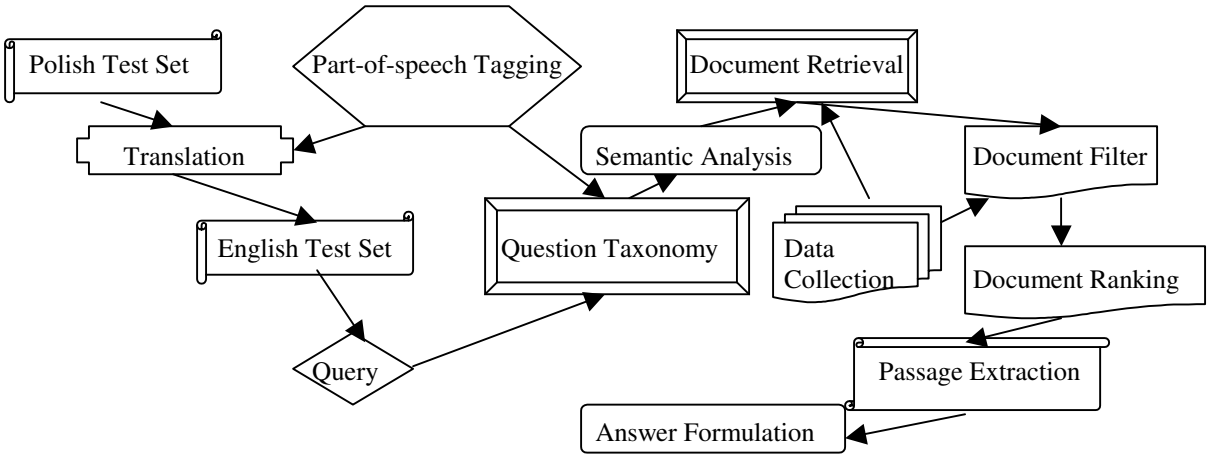

Figure 1. The Architecture of the system AID

2.1 Direct MT Model

AID was submitted a Polish test set consisting of two hundred questions of the following types: factoid questions, definition questions, temporary restricted by date, period or event questions and list questions. As a tool LEC automatic machine translation software by Language Engineering Company LLC , USA was utilized. It employs three kinds of dictionaries: specialized, universal (Collins) and technical ones. Actually it can translate 272 language pairs. The software consists of several applications: LEC ClipTrans, LEC Dictionary, LEC FileTrans, LEC LogoTrans, LEC MirrorTrans, LEC Transit and LEC Translate. This way a user can obtain either a batch translation of a file, Web page, e-mail or MS Office Pack automatically.

The company announces that all the translations are direct which means the output is organized in the target language sentence format [see FEMI for details]. Before performing word-for-word or phrase-for-phrase translation some simple morphological sentence analysis aimed at word reordering based on the part-of-speech tagging is implemented. However it is a very straight forward technique, it proves quite efficient in question answering because of the limitation of the grammar structures in the test set.

All of the questions employ either Simple Present or Simple Past tenses and occasionally Present Perfect. Consequently, the words are marked up according to their position in the question. In the next stage each of the words is translated automatically by the inbuilt dictionary.

Figure 2 shows the translation scheme on the base of the test set submitted to AID.

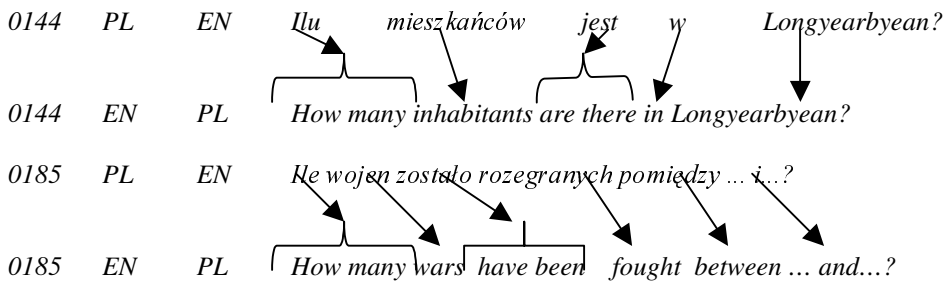

Figure 2. Word alignments in the Direct Translation model

Sometimes more than one word have the same translation. In the example above both Polish "Ile" and "Ilu" have been translated into "How many". In sentence 0185 the correct translation of "zostało" requires using the Present Perfect tense "have been". These two examples represent linguistic phenomena between the language pair.

2.2 English Test Set Processing

In this section, the process of the query analysis is described. At first, the question taxonomy relies on question classes and subclasses. The test set comprises nine question classes: WHAT, WHICH,WHO, WHERE, WHEN, HOW MANY, HOW MUCH, NAME and LIST. Subclasses on the other hand, constitute the questions that include preposition or a subject associated with the question word. Eventually, both a preposition and a subject.
The system component Semantic Analysis preprocesses the query by extracting some keywords that determine the type of the document to be indexed and then the text portion relevant to the information required. Query reformulation has been abandoned for a simple reason – in natural interactive verbal communication it is impossible to predict the form of the answer. In this light, any number of suggestive answer forms cannot result with the better system performance what is confirmed by the research reports [e.g. Brill E. et al., 2003].
AID collects all the documents in which the keywords occur. The only fields analyzed in the document collected by the system are: document number, date and text.

```
<DOC>
<DOCNO>GH951127-000092</DOCNO>
<DOCID>GH951127-000092</DOCID>
<DATE>951127</DATE>
<HEADLINE>Worth a thought</HEADLINE>
<EDITION>1</EDITION>
<PAGE>11</PAGE>
<ARTICLETYPE>Correspondence</ARTICLETYPE>
<FLAG>sport</FLAG>
<RECORDNO>975175773</RECORDNO>
<TEXT>
THE confusing incident towards the end of the match between Scotland and
Western Samoa indicates the need for a new rule about which I have written
previously. Again, I propose that penalty kicks for illegal offences …
Dr John A J Macleod,
Western Isles.
</TEXT>
</DOC>
```

Figure 3. A standard format of the documents in British collection.

As Figure 3 presents, data processing was relatively easy for the system because all the documents have the same format on the contrary to the Web documents. In other words, we can call this type of collection as homogeneous.

The technique relying on eliminating entities that do not influence the information quality seems essential in case of large collections. This document format consists of many entities that are unnecessary in respect to the information relevance.

2.3 Information extraction

For document selection AID deploys so called shallow technique that ranks the documents according to the syntactic similarity of the snippet to the question. The position of the keywords in the question and the answer determine reliability of the rank criteria. Lin's algorithm [Lin et al., 2003] used at this stage can be expressed as follows:

$$\frac{1}{|A|}\sum_{w\in A}\log(\frac{N}{w_c})$$

where:
A is a set of keywords ,
N total number of words in the corpus,
w_c number of occurrences of word ω in the proposed answer.
The system scores the relevant answers aligning the query keywords, the question word and the question notion according to the question taxonomy (classes and subclasses). The first document awards the highest score and the last ranked within the corpus, the lowest one. The most relevant documents seem to be those that retrieve the answers with the greatest number of the words associated both in the query and the snippet supporting the answer.
The approach can be illustrated by the following example of definition question type:

What is Atlantis?

The question word is WHAT, the keyword is ATLANTIS and the notion is NAME.. The system response to this question was:

```
R     0001  utjp061plen 0.981 LA110794-0104     space shuttle     the space
shuttle Atlantis
```

The answer was supported by "the space shuttle Atlantis" so the words between "the" and "Atlantis" form a subject that is why the confidence score is close to 1. Again, instead of grammatical analysis of the whole texts or the documents, AID focuses only on the words associated with the keywords. This methodology constitutes the backbone of the AID system.

3 The Experiment Framework

Despite registering for six out of eight tracks, our group submitted only one run named utjp061plen for Polish-English question answering task. This work has been proceeded by translation of the 200 questions into Polish and earlier the test set for the Ad Hoc track. Both translations have been posted on the CLEF Web site to be available for other participating groups.
AID has been tested on two data collections:

- Los Angeles Times 1994 that contains 113,005 American documents and requires 425 MB
- Glasgow Heralds 1995 that contains 56,472 British documents and requires 154 MB of HDD.

Additionally, for previously planned Polish-French Ad Hoc track four other data collections have been downloaded:

- Le Monde 1994 that contains 44,013 French documents and requires 157 MB
- Le Monde 1995 that contains 47,646 French documents and requires 156 MB

- SDA French 1994 with 43,178 documents that require 86MB
- SDA French 1995 with 42,615 documents that require 88 MB

As intended at the beginning these six document collections have been downloaded on computer with Intel Celleron 370 MHz processor and 128 MB RAM. The process took around eight hours. Thus, the experiment was carried out on another computer with Pentium 1,6 GHz processor and 512 MB DDR2 working on two platforms: Windows XP and Linux.

For decompression of the databases the gzip program was used whereas Jade DSSSL engine with backend that generates RTF formats empowered printing and displaying SGML documents. Other tools used in this experiment have been mentioned in the previous sections.

4 Analysis of the System Performance

The system performance was evaluated manually on the base of responsiveness and exactness. The following judgments apply to the procedure: right (R), inexact(X when the answer is not complete), unsupported (U when the snippet is incorrect), wrong (W), and not assessed (Z).

Table 1 shows the judgments in relation to the question classes.

Question Class	R	W	U	X	Total number	Accuracy [%]
WHAT	63	5	5	2	75	84
WHICH	41	2	1	0	44	95
WHO	32	2	1	1	36	94
WHERE	10	2	0	0	12	83
WHEN	6	2	0	0	8	75
HOW MANY	17	0	0	0	17	100
HOW MUCH	2	0	0	0	2	100
NAME	2	0	4	0	4	50
LIST	2	0	0	0	2	100

Table 1. Accuracy of AID performance measured over the question classes.

This table presents the system accuracy within the class taxonomy. The overall accuracy of the right answers was as high as 86.32%. Such a result is very promising . However the table indicates the impact of the number of some question classes on the overall result. Out of two hundred questions AID produced 164 right answers, 14 wrong ones, 5 inexact ones, 7 unsupported and 0 not assessed. Assuming that the questions are dealt into definition, factoid, temporary restricted and list questions its accuracy equals accordingly to 88%, 80% and 0%. Instead of retrieving answer NIL 25 times, AID responded correctly in this case only 17 times which is 68%.

Regarding the list questions, out of 31 correct answers, AID produced 18 right answers (58%), 0 wrong ones, 11 unsupported, 2 inexact and again 0 not assessed. This question type proved the most difficult for AID.

As a result, the precision P@N = M / N where M is the number of right responses and N is the number of answers judged per question is 0.65 for the list questions.
For factoid and definition questions the evaluation measure was based on the MRR (Mean Reciprocal Rank) used also in TREC evaluation. It represents the mean between 0 in case of no correct responses and 1 when the system produces all the right answers at position 1.
The overall Confidence Weighted Score for the system performance is CVS 151.602/190 = 0.79790

4.1 The System Strong Sides

The overall score for the system performance proved excellent. This has been achieved by limitation of fields retrieved at the Document Retrieval stage, direct machine translation model used by the LEC tool and predominantly by methodology applied for the limited information extraction. Furthermore, AID benefits from its modularity and simplicity.

4.2 The System Limitations

It is equally important to mention that such a great accuracy lies in environment which is meant as specified conditions. AID has been built and adopted to the CLEF campaign requirements. It was rather a fast forward approach with its limitations being a consequence of the experiment deadlines.
Thus, the first drawback is that we can talk about information extraction rather than answer formulation. Then, as yet AID was tested only on homogeneous collections not for instance on the net. The question taxonomy includes only nine question classes. With regard to the list questions, when the passage contains the target answer of two or more sentences, or the keywords are replaced by their synonyms AID fails producing NIL as an answer.
The list questions are processed correctly but on condition that the snippet contains the words or phrases associated. In any other case the names that are in isolation remain missed by the system.
And eventually, the score for overall system accuracy depends on the formula that represents number of definition or factoid questions to the list or temporary restricted questions.

5 Conclusion

This work is found by me as a great achievement for a number of reasons. The first one is that testing so many question answering systems gave me an idea about the AID methodology and the detailed concept of its architecture. It imposed on me browsing the Internet for the tools, sharing experience with the researchers almost all over the world working in the field, studying unknown yet software instructions in terms of having an intensive course in specific areas of information retrieval.
As for the system performance, it gained a very high accuracy score in comparison to other participating groups and Polish was used for the first time in the CLEF campaign. The section above which describes the system limitations indicates how much work is to be done in the future so as to expand the system capacities making it to become a real open-domain question answering Polish-English system.

Acknowledgements

I am grateful to my Adviser, Professor Aleksander Zgrzywa and all the colleagues from my university for their invaluable comments. Without them this work would not be completed on time and especially the accuracy of the system performance would not achieve that high score. I would like to thank a student Jakub Felski for sharing with me His knowledge and experience in the field, as well as being always on hand. I want to thank Ms Jozefa Bernardyn for Her time and presentation of Knoppix.
I would like to express my gratitude to Professor Felisa Verdejo and the organizers for their kindness and support with my participation in the conference events. I also want to thank dr. Carol Peters for arranging the enterprise aimed at integrating our research community. My participation in the Doctoral Consortium and the CLEF campaign gave a final shape to my Ph. D. thesis.

References

Brill E., Dumais S., Bank M.: *An Analysis of the AskMSR Question Answering System*, Microsoft Research, One Microsoft Way, 2003.

FEMI – *a Framework for the Evaluation of Machine Translation in ISLE*, Information Science Institute, USC Viberti School of Engineering, http://www.isi.edu/natural-language/mteval

Guzman R: *LogoMedia's Translate – What does it translate*?, Localization Ireland, No 10, 2001.

Lin J., Katz B.: *Question Answering from the Web Using Knowledge Mining Techniques*, Proceedings of the 12th International Conference of Information and Knowledge Management, 2003.

Strötgen R, Mandl T., Schneider R.: *A Fast Forward Approach to Cross-lingual Question Answering for English and German*, Working Notes, CLEF 2005.