QUESTION ANSWERING AS A KNOWLEDGE DISCOVERY TECHNIQUE

Yllias Chali
University of Lethbridge, 4401 University Drive
Lethbridge, Alberta, Canada, T1K 3M4

Keywords: Knowledge and Information Extraction, Question Answering.

Abstract: The size of the publicly indexable world-wide-web has provably surpassed several billions of documents and as yet growth shows no sign of leveling off. Search engines are therefore increasingly challenged when trying to maintain current indices using exhaustive crawling. Focused Retrieval provides more direct access to relevant information. In this paper we investigate the various aspects of discovering knowledge about entities such as people, places, groups, and about complex events.

1 INTRODUCTION

The size of the publicly indexable world-wide-web has provably surpassed several billions of documents and as yet growth shows no sign of leveling off. Dynamic content on the web is also growing as time-sensitive materials, such as news, financial data, entertainment and schedules become widely disseminated via the web. Search engines are therefore increasingly challenged when trying to maintain current indices using exhaustive crawling. Even using state of the art systems such as AltaVista's Scooter, which reportedly crawls ten million pages per day, an exhaustive crawl of the web can take weeks. This vast raise in the amount of online text available and the demand for access to different types of information have, however, led to a renewed interest in a broad range of Information Retrieval (IR) related areas that go beyond simple document retrieval, such as focused retrieval, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), software engineering, chemical and biological informatics, text structuring, text mining, and genomics (Voorhees, 2003a; Voorhees, 2003b). Focused Retrieval (FR) is relatively a new area of research which deals with retrieving specific information (i.e. passage or answer to a question or XML element) to the query rather than state of the art information retrieval systems (search engines), which retrieve documents (Harabagiu et al., 2003; Moldovan et al., 1999; Roth et al., 2002; Moldovan et al., 2002). This means that the focused retrieval systems will possibly be the next generation of search engines. What is left to be done to allow the focused retrieval systems to be the next generation of search engines? The answer is higher accuracy and efficient extraction. In this paper, we investigate various aspects of the focused retrieval applications such as question answering, passage retrieval and element retrieval. We are proposing in this paper techniques to extract useful information about entities such as people, places and groups, and about complex events. To achieve this goal, we need to develop mechanisms to answer questions about these entities and complex events. For instance, considering "Abraham Lincoln", we can extract the information that are answers to the following questions:

Who is Abraham Lincoln?
When was he president?
When did he die?
How did he die?
Who shot him?
etc.

This paper is organized as follows. Section 2 presents the pre-processing techniques consisting of the normalization of the questions. Then, we describe our system for question answering in all its details. Finally, we present an evaluation of the system and conclude by some future works.

Chali Y. (2008).
QUESTION ANSWERING AS A KNOWLEDGE DISCOVERY TECHNIQUE.
In *Proceedings of the Fourth International Conference on Web Information Systems and Technologies*, pages 339-342
DOI: 10.5220/0001531003390342

2 QUESTION NORMALIZATION

The questions are not only about entities but could be about complex events such as "the visit of Prince Charles and Camilla to California". We call the theme of the questions in general the *target*. The questions are grouped by target being the overall theme of the questions. The targets are mainly people, places, groups, and events. For instance, we could have a target like "Space Shuttles," and we will have all possible questions about this target or more specific questions about more topics like "Spaceship Columbia". Our goal is to collect and discover several useful information about a specif target. To accomplish this goal, we need to answer all the possible questions about the target. For instance, if we consider Canada general election event as a target, we will have the scenario including the following questions:

Target:
Canada general election
Questions:
When was the last Canada general election?
Why was the election called?
Which political parties participated
in the election?
Who was leading the liberal party?
Who was leading the conservative party?
What was the election results?
How many seats did each party get
in the parliament house?
What are the changes with the new
government?
How long is the Canadian mandate?
etc.

The question normalization module takes the information given by the target and the questions, and changes the questions to incorporate that information. This means that questions can refer to the target of the questions, or even to other questions. Our system resolves these references so that it can answer the questions one at a time.

These types of references were also investigated by (Schone et al., 2004). We classify these references in three ways:

- Reference to the target by a pronoun
- Reference to the target by an entity
- Implied Reference

Our system resolves the pronouns of the question first, then proceeds to resolve the other entities.

2.1 Pronoun Resolution

Our system assumes that pronouns are referring to the target of the question, unless the target already appears in the question. It considers two types of pronouns; personal and possessive.

Personal pronouns just need a direct replacement with the target. An example of this is the question, "Where was he born?", with the target, "Walter Mosley", which will be changed to "Where was Walter Mosley born?" The personal pronouns we are considering are; it, he, she, they, him and her.

The possessive pronouns will involve more than a direct replacement. An "'s" will be added after the target, once the target replaces the pronoun. An example of this is the question, "Who is her coach?", for the target, "Jennifer Capriati", which will be changed to "Who is Jennifer Capriati's coach?"

2.2 Entity Resolution

These are references to the target, or a past question, in the form of what type of entity it is. An example of this is the entity, "the cult", in the question, "Who was the leader of the cult?". This entity is referring to the target, "Heaven's Gate". These entities start with "the" or "this", and they could refer to three things; the target, an answer from a previous question, or the answer to the question.

First our system checks for a pattern correlation between the entity and the target. For instance the question, "When was the agreement made?", has the entity "the agreement" which corresponds to the target "Good Friday Agreement", so the entity will be replaced by the target.

Each time an entity represents an answer, the entity is saved along with its corresponding answer. Then, if that entity appears again, it will be replaced with the answer, if it has not already been replaced by the target. For the question, "What are titles of the group's releases?", the entity that needs resolution is "group", which does not correspond to the target "Fred Durst". It does, however, correspond to the focus of a previous question, "What is the name of Durst's group?" Therefore, the entity "group" from the question, "What are titles of the group's releases?", will be replaced by the answer of the previous question.

2.3 Implied References

Implied references are when the target is implied, but not explicitly stated. An example of this is the question, "Who was President of the United States at the time?", for the target, "Teapot Dome scandal". The question would ideally be reformed to "Who was President of the United States at the time of the Teapot Dome scandal?", but our system does not reform the question in such a way. Our system will include the target in the query for these questions without reforming the question itself.

Some questions are more difficult than this and need to be treated differently. The question, "How many are there now?", is looking for a count of an entity which is the target, which is "Kibbutz" in this case. Therefore, if our system can not determine the entity that is to be counted, it will consider the target as the entity.

3 OUR SYSTEM

Once the questions are normalized, each question is answered individually by our question answering system. Figure 1 outlines the general architecture of our system. In the subsequent, we detail each of its components.

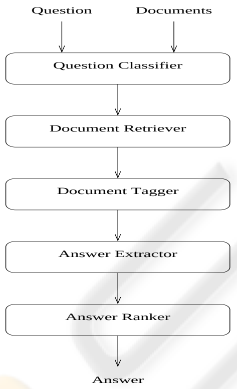

Figure 1: System Overview.

3.1 Question Classification

Questions are classified by first separating them into one of the following categories; *who*, *when*, *why*, *how*, *where* and *what*. If a question is not easily classified as one of the above, it will be classified as a *what* question.

After they are categorized, the named entity (NE) the answer will take is found. For *what* questions this may involve discovering the question focus. The focus of a question is the part of a question that tells what type of entity the answer will be. For instance, the focus of "What city is home of the CN Tower?" has a focus of "city". We use a group of patterns to discover the focus in *what* questions.

3.2 Document Retrieval

We are using Managing Gigabytes (MG) (Witten et al., 1999) for our information retrieval system. We separate each document into paragraphs, and index each paragraph as if it were a document. When a question is being processed, the question classification module creates a boolean query for MG to retrieve the documents.

3.3 Document Tagging

We use Collins Parser (Collins, 1996) and OAK Tagger (Sekine, 2002) to tag the documents that are retrieved by MG. Collins Parser tags the word dependencies from the documents, and the OAK Tagger tags chunked parts of speech and named entities that correspond to the answer type of the question. The documents that are parsed, and the documents tagged by OAK tagger, will be sent to the answer extractor.

3.4 Answer Extraction

The two sets of tagged documents will have their answers extracted differently.

For the parsed set of documents, the question parse tree will be used to fill in the missing information from the parse tree of the documents. If an entity can be found such that it can complete the answer parse tree, it is passed to the answer ranker.

For the OAK tagged documents, if they contain a named entity corresponding to the answer type of the question, they are then passed to the answer ranker.

For some other questions, patterns for both parsed documents and the chunked part of speech documents will extract possible answers to be ranked by the answer ranker module.

3.5 Answer Ranking

If the answer type of the question corresponds to a tagged named entity, all the entities extracted from the tagged documents will be considered possible answers. They will be ranked by how many times they appear in the possible answer list, how close they appear to words from the question, and if they appear in the list of entities extracted from the parsed documents. If the answer type is not a named entity, the entities extracted from the parsed documents will be considered possible answers and will be ranked only on frequency.

For factoid questions, the top ranked possible answer is given as the answer to the question if it achieves a rank above the threshold for the type of question. For list questions, the possible answers that achieve a rank higher than the threshold will be given as the answer. For other questions, possible answers that appeared more then two times are given as answers. This is because our patterns sometimes extract useless information, and if a piece of information is important about a target, it will usually get extracted more than once. A useless fact should only be extracted once from the set of documents.

4 EVALUATION

The TREC question answering track provides the testing data to evaluate the accuracy of the systems. It consists of sets of documents and questions/answers related to those sets of documents. We evaluate our system considering these data. The results of the evaluation are shown in Table 1.

Our system still not ready for all the types of questions that are asked in the TREC QA track collection. This difficulty arose because we mainly train our system on the questions and answers, and we do not present a corpus of questions large enough to include classifications for the questions. Therefore, we lose in the accuracy of attempts to answer questions.

Table 1: Evaluation Results shown by Question Categories.

Question Type	Success
Who	0.317
When	0.328
Why	0.245
How	0.265
Where	0.345
What	0.294
List	0.308
Others	0.145
Overall	0.281

5 CONCLUSIONS

We presented a system that extracts information about entities and events given a pool of questions related to that entity or event. We create categories for all the questions. We extract rules to classify questions into each of these categories. The system also includes syntactic features and part of speech features for the question classification and answer extraction.

Our system still need some improvements. The overall improvement is primarily expected by the expanded classification of questions and the addition of dependency features to answer finding (Li and Roth, 2005; Pinchak and Lin, 2006). We hope to carry on this research and obtain an even greater improvement.

REFERENCES

Chali, Y. and Dubien, S. (2004). University of Lethbridge's participation in TREC-2004 QA track. In *Proceedings of the Thirteenth Text REtrieval Conference*.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of ACL-96*, pages 184–191, Copenhagen, Denmark.

Harabagiu, S., Moldovan, D., Clark, C., Bowden, M., Williams, J., and Bensley, J. (2003). Answer mining by combining extraction techniques with abductive reasoning. In *Proceedings of the Twelfth Text REtrieval Conference*, pages 375–382.

Li, X. and Roth, D. (2005). Learning question classifiers: The role of semantic information. *Journal of Natural Language Engineering*.

Moldovan, D., Harabagiu, S., Girju, R., Morarescu, P., Lactusu, F., Novischi, A., Badulescu, A., and Bolohan, O. (2002). LCC tools for question answering. In *Proceedings of the Eleventh Text REtrieval Conference*.

Moldovan, D., Harabagiu, S., Pasca, M., Mihalcea, R., Goodrum, R., Girju, R., and Rus, V. (1999). LASSO: A toll for surfing the answer net. In *Proceedings of the 8th Text REtrieval Conference*.

Pinchak, C. and Lin, D. (2006). A probabilistic answer type model. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 393 – 400.

Roth, D., Cumby, C., Li, X., Morie, P., Nagarajan, R., Rizzolo, N., Small, K., and Yih, W. (2002). Question-answering via enhanced understanding of questions. In *Proceedings of the Eleventh Text REtrieval Conference*.

Schone, P., Ciany, G., P. McNamee, J. M., Bassi, T., and Kulman, A. (2004). Question answering with QACTIS at TREC-2004. In *Proceedings of the Thirteenth Text REtreival Conference*.

Sekine, S. (2002). Proteus project oak system (English sentence analyzer), http://nlp.nyu.edu/oak.

Voorhees, E. M. (2003a). Overview of the TREC 2002 Question Answering track. In *Proceedings of the Eleventh Text REtrieval Conference*.

Voorhees, E. M. (2003b). Overview of the TREC 2003 Question Answering track. In *Proceedings of the Twelfth Text REtrieval Conference*.

Witten, I., Muffat, A., and Bell, T. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann.