A Statistical Language Modelling Approach for Question Answering System

Mr. M. Arutselvan[*], Mr. T. Chellatamilan[**], Mr. M. Vijaya Kumar[**]

[*] Department of Computer Science and Engineering, Arunai Engineering College Tiruvannamalai, India
[**] Department of Computer Science and Engineering, Arunai Engineering College, Tiruvannamalai, India
[***] Unifying Logics Pvt Ltd, Software Engineer Coimbatore, Tamil Nadu, India

Abstract- This paper concerns about Question Answering system in which a statistical language modeling approach is used. The main objective is to build a simple system for question answering without the need for highly tuned linguistic modules which need more human work and is very difficult to find any bugs if any. A mathematical model for answer retrieval and answer extraction is derived, which does not use any linguistic information or annotated data. It makes use of word tokens and web data. We take a statistical, noisy-channel approach and treat QA as a whole as a classification problem. We present a fully data-driven mathematical model for estimating the probability of a candidate answer given a question. In doing so we largely remove the need for ad-hoc weights and parameters that were a feature of many TREC systems.

Index Terms- Question Answering, Linguistic Modules.

I. INTRODUCTION

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. An information retrieval process begins when a user enters a query into the system. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a database. Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. Question Answering (QA) concerns itself with the development of systems that can automatically and accurately answer questions posed in natural language, and draws upon fields such as information retrieval (IR), natural language processing (NLP) and machine learning Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. Many challenges in NLP involve natural language understanding that is, enabling computers to derive meaning from human or natural language input.

Modern NLP algorithms are based on machine learning, especially statistical machine learning. An early example is Murax an open-domain QA system which combined robust linguistic methods with an IR system in order to find answers in an online encyclopedia.

II. METHODOLOGY

NLP techniques employed by QA systems typically include part-of-speech (POS) tagging, named entity (NE) extraction, parsing and query expansion. The best performing systems in TREC evaluations have become increasingly complex, relying on a number of modules using highly-tuned, sophisticated NLP techniques, usually with great manual effort. There have also been many attempts to diverge from the complex linguistic approaches towards more robust, data-driven approaches, exploiting the huge domain coverage and redundancy inherent in web data. Redundancy in web data may be seen as effecting data expansion, as opposed to query expansion techniques and complex linguistic analysis often necessary in answering questions using a small corpus, such as the AQUAINT corpus, containing around 1 million documents.

III. LITERATURE SURVEY

As the symbolic approaches to NLP gave way to more empirically driven research in the 1990s, open-domain QA systems were developed, which relied on more shallow linguistic processing and IR on unstructured data corpora. The availability of large amounts of data, both for system training and answer extraction, logically leads to examining statistical approaches to QA. Several non- linguistic, statistical methods were investigated for what was termed bridging the lexical gap between questions and answers, such as maximum-entropy based query expansion, as well as statistical translation models where the question is considered the source language and the answer the target language.

A statistical translation model is also used in [4] to bridge the lexical gap, but extends the previous mentioned work by formulating the answer extraction problem in terms of a noisy channel model. In [4] a maximum-entropy based classifier using several different features was used to classify answers as correct or incorrect. A statistical noisy-channel model was used in [6] in which the distance computation between the query and the candidate answer sentences is performed in the space of parse trees.

In this paper we present a QA approach which, of the mentioned works, is most similar to [4] and the re-ranker. We take a statistical, noisy-channel approach and treat QA as a whole as a classification problem. We present a fully data-driven mathematical model for estimating the probability of a candidate answer given a question. In doing so we largely remove the need

for ad-hoc weights and parameters that were a feature of many TREC systems. Our motivation is the rapid development of data-driven QA systems in new languages where the need for highly tuned linguistic modules is removed. Apart from our mathematical model for QA, the main difference between our approach and many contemporary approaches to QA is that we only use word tokens in our system and do not employ NE extraction or any other linguistic information, e.g. from semantic analysis or question parsing; nor do we use hand-crafted or annotated lexical resources such as WordNet.

IV. SYSTEM ARCHITECTURE

A typical state-of-the-art QA system architecture has a question analysis module which processes a question posed by the user. It constructs a query that is used by an IR module, as well as answer type information that is used by an AE module. The IR module retrieves documents or passages from a corpus, e.g. a newspaper corpus or the World Wide Web, and passes them to the AE module. Ideally the retrieved question and answer type information, answer hypotheses are extracted and presented to the user as a ranked list. The architecture of our QA

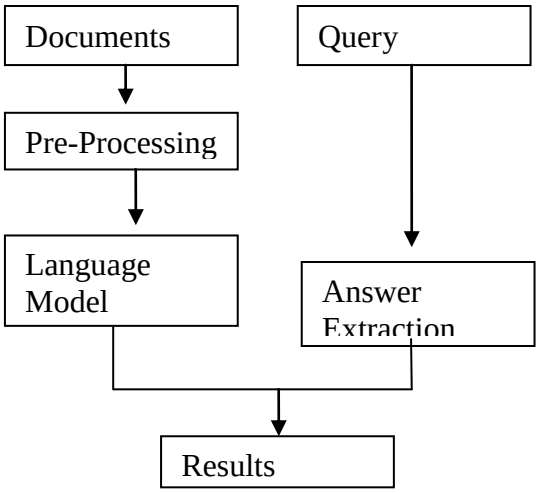

Fig.1 Architecture of our QA system

System is shown in Fig. 1 follows this pattern closely, although it doesn't have a separate question analysis module. Question processing is an integral part of the IR and AE modules and only involves tokenization and removal of stop-words. Moreover, answer type analysis is not explicitly performed, but implicitly done in the AE module.

In the AE module candidate answers are extracted from the retrieved text. These candidate answers are then ranked according to the probability of the candidate answer *A* given the question *Q*. The next section explains in detail how this probability is estimated.

V. PREPROCESSING

In this phase of the Project, document collections are pre-processed in order to remove stopwords and to store other words in dictionary. This process is repeated for queries also. A stop list which contains a minimum of 500 words is used for pre-processing the document collections and queries. A dictionary is maintained for storing and updating the number of words which are not stopwords, for further processing.

VI. LANGUAGE MODEL GENERATION

The Statistical Language Model mainly concerns on a probabilistic distribution over a word sequences. It provides a principled way to quantify the uncertainties associated with the natural language. After the pre-processing process language model is generated for each and every document. That is for every words present in a document a probabilistic value is assigned.

$$P(t|D) = P(t|\theta_D) = \frac{tf_{t,D}}{N_D} \quad (1)$$

VII. ANSWER EXTRACTION

It is the most important phase of the project in which the documents which are relevant to the given query will be extracted from the document collections. This process is done by generating language model for both query and document collections and the value of query is compared against every document. The document whose value is higher than any other document will be ranked as one and other documents are ranked vice versa..

We consider the dependence of an answer *A* on the question *Q*, where each is considered to be a string of $|A|$ words $A = (a1 , .. ., a|A|)$ and $|Q|$ words $Q = (q1 , .. ., q|Q|)$, respectively. In particular, we hypothesize that the answer *A* depends on two sets of features $W = W(Q)$ and $X = X(Q)$ as follows:

$$P(A|Q) = P(A|W, X),$$

where $W = \{w1 ,..., w|W |\}$ can be thought of as a set of $|W|$ features describing the "question-type" part of *Q* such as *when*, *why*, *how*, etc. and $X = \{x1 , .. ., x|X| \}$ is a set of $|X|$ features comprising the "information-bearing" part of *Q*, i.e. what the question is actually about and what it refers to.

$$Score(Q, D) = P(Q, \theta_D) = \prod_{i=1}^{n} P(q_i|\theta_D)$$

VIII. EXPERIMENT

For our experiments we use the factoid questions from the TREC QA tracks. Text processing of corpus, web data, questions and answers is intentionally minimal; it involves only removing unnecessary mark-up and CACM datasets for searching documents. By using the query sets documents are ranked and listed. In our experiment, ranked documents are listed. The precision and recall values are calculated for the listed documents and a graph is drawn based on these values.

	D1	D2	D3	..	D10
Precision	0.0065	0.004	0.0023		0.0001
Recall	0.04	0.06	0.08		1

IX. CONCLUSION

In this paper we have presented a statistical, fully data-driven Question Answering system. A mathematical model for Answer Extraction was derived, which estimates P(A|Q), the probability of an answer candidate A given a question Q, and is decoupled into two independent models: a retrieval model and a filter model.

REFERENCES

[1] Whittaker, E., Furui, S., Klakow, D., 2005. A statistical classification approach to question answering using web data. In: Proceedings of the 2005 International Conference on Cyberworlds, pp. 421–428.

[2] Whittaker, E., Novak, J., Chatain, P., Furui, S., 2006. TREC2006 question answering experiments at Tokyo Institute of Technology. In: Proceedings of the 15th Text Retrieval Conference (TREC 2006).

[3] Ittycheriah, A., Roukos, S., 2002. IBM's statistical question answering system. In: Proceedings of the 11th Text Retrieval Conference (TREC 2002).

[4] Soricut, R., Brill, E., 2004. Automatic question answering: beyond the factoid. Ponte, J.M., Croft, W.B., 1998. A language modeling approach to information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 275–281.

[5] Brill, E., Dumais, S., Banko, M., 2002. An analysis of the AskMSR question-answering system. In: Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 257–264.

AUTHORS

First Author – M. ArutSelvan.,M.E., Arunai Engineering College, sanarut8@gmail.com.
Second Author – Mr. T ChellaTamilan Arunai Engineering College,chellatamilan_t@gmail.com.
Third Author – M. VijayaKumar Unifying Logics Pvt Ltd,viji5684@gmail.com.