IBM's Statistical Question Answering System – TREC-11

Abraham Ittycheriah, Salim Roukos

IBM T. J. Watson Research Center
P.O.Box 218
Yorktown Heights, NY 10598
{abei,roukos}@us.ibm.com

Abstract

In this paper, we document our efforts to extend our statistical question answering system for TREC-11. We incorporated a web search feature, and novel extensions of statistical machine translation as well as extracting lexical patterns for exact answers from a supervised corpus. Without modification to our base set of thirty-one categories, we were able to achieve a confidence weighted score of 0.455 and an accuracy of 29%. We improved our model on selecting exact answers by insisting on exact answers in the training corpus and this resulted in a 7% gain on TREC-11 but a much larger gain of 46% on TREC-10.

1 Introduction

TREC evaluations in Question Answering provide a useful application benchmark, which allows validation of a number of component technologies for which evaluation criteria are absent by providing a score for the integration of these components. Our approach since TREC-9 has been to investigate a mathematical framework under which a useful solution for question answering could be produced. We will present our model and its novel extensions below. For training our system, we collected a 4K question-answer corpus based on trivia questions and developed answer patterns for the TREC collection of documents. This corpus was used to drive a number of components we will describe below. This corpus also allowed us to investigate weights on features such as presence of the answer chunk in web documents and lexical patterns found in answers. We also describe our efforts after the evaluation to overcome the inexact answer problem and present results obtained since the evaluation.

In TREC-8 (Voorhees and Tice, 1999), the NLP community began the task of evaluating Question Answering systems and has in subsequent evaluations provided significant challenges to such systems. In TREC-9, the challenge was 50-byte answers and in TREC-10 it was definitional questions and handling rejection. To address these challenges, systems have largely adopted the architecture of predicting the answer tag of the desired answer, using a document retrieval method to select relevant documents and performing answer selection to obtain the target answer. In TREC-8 (Srihari and Li, 1999) obtained significant gains using an expanded class of entities (66). In TREC-9, improved performance was demonstrated by using boolean retrieval and feedback loops (Harabagiu and et. al., 2000). In TREC-10, use of a large number of patterns was shown to perform well for retrieving answers (Soubbotin, 2001). In TREC-11, the track agreed to several significant changes

- Exact Answers
- Single Answers
- Confidence-based Ranking of Answers

1.1 Exact Answers

Systems were required to return answers which had only the desired answer. Extra words were not accepted and their presence caused the answer to be judged inexact. Our approach of handling exact answers was to use our phrases spanned by our thirty named entity categories as well as constituent phrases of the syntactic parse of the answer (Penn Treebank style) which satisfied the answer pattern for the question. The decision to use the syntactic parse based phrases caused our system to output a large number of answers which were judged as inexact. We will describe some experiments where we changed the decision to accept only those phrases which exactly satisfy the answer pattern. Our named entity categories do not capture the differences between dates and years; nevertheless, we decided to evaluate our system without modifying the named entity categories. The named entity tags are broken along five major categories:

Name Expressions Person, Salutation, Organization, Location, Country, Product

Time Expressions Date, Date-Reference, Time

Number Expressions Percent, Money, Cardinal, Ordinal, Age, Measure, Duration

Earth Entities Geological Objects, Areas, Weather, Plant, Animal, Substance, Attraction

Human Entities Events, Organ, Disease, Occupation, Title-of-work, Law, People, Company-roles

1.2 Single Answers

In previous TREC evaluations, systems returned upto 5 answers per questions. In TREC-11, only a single answer was returned for each question. For evaluating single answers, the criteria used in this evaluation was the accuracy of the system.

1.3 Confidence-based Ranking of Answers

NIST changed the metric from the mean reciprocal rank (MRR) of previous TREC Q&A evaluations to the Uninterpolated Mean Average Precision, which we shall refer to as the confidence weighted score (CWS) defined as follows,

$$\text{CWS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\#\text{ correct upto question } i}{i}$$

where N is the number of questions. This metric gives more credit to questions answered correctly at the beginning of the list. We made no specific attempt to optimize on this criteria and instead worked mostly on optimizing the accuracy of our system.

2 TREC 11 System

We model the distribution $p(c|a,q)$, which attempts to measure the c, 'correctness', of the answer and question. c can take on values of either 0 and 1 indicating either an incorrect or correct answer respectively. We introduce a hidden variable representing the class of the answer, e, (answer tag/named entity) as follows,

$$\begin{aligned} p(c|q,a) &= \textstyle\sum_e p(c,e|q,a) \\ &= \textstyle\sum_e p(c|e,q,a)p(e|q,a) \end{aligned} \tag{1}$$

The terms, $p(e|q,a)$ and $p(c|e,q,a)$ are the familiar answer tag problem and the answer selection problem. Instead of summing over all entities, as a first approximation we only consider the top entity predicted by the answer tag model and then find the answer that maximizes $p(c|e,q,a)$.

The distribution $p(c|e,q,a)$ is modeled utilizing the maximum entropy framework described in (Berger et al., 1996). We built on top of the model we used last year and those features are described in (Ittycheriah et al., 2001). The new features we investigated for this year are:

- Occurrence of the answer candidate on the web
- Re-ranking of answer candidate window using a statistical MT dictionary
- Lexical patterns from supervised training pairs

This year we submitted 3 runs, two of which measured the effectiveness of the first feature type. The last run was a feedback loop on the first run, where we included the answer string of questions which had sufficient confidence to further improve their confidence. The results are presented below in Table 1. We also provided the output of system 'ibmsqa02a' to another group at IBM for the run labeled IBM-PQSQA. The integration of our system's output with their question answering system, improved their base performance from 33.8% to 35.6%, an improvement of 5.3% in accuracy and in terms of CWS, from 0.534 to 0.586 (Chu-Carroll et al., 2002).

2.1 Training Data

We used TREC-8, TREC-9, and 4K questions from our KM database to train the model this year. This corpus represented an order of magnitude increase in size over the training data size we used last year. For each question we developed a set of answer patterns by judging several potential answer sentences in the TREC corpus. Using the answer patterns and sentences derived from the TREC corpus, we automatically labelled chunks as being correct or incorrect. The total number of chunks used in formulating the model was 207K. There were 30K instances of correct answers (though 10K were inexact) and 177K incorrect chunks.

2.2 Web Feature

The web feature was used by a number of groups last year (Clarke et al., 2001) (Brill et al., 2001) and we attempted to measure its impact on our system. We incorporated the feature as two indicators: (1) occurrence of the answer candidate in the top 10 documents retrieved from the web, (2) count of the number times the answer candidate occurred. This feature type and performing no rejection is the difference between the runs ibmsqa02a and ibmsqa02b. Removing rejection the correctly rejected questions, we note only an improvement of 7 questions by using the web based feature. Our systems have traditionally used an encyclopaedia for LCA based expansion and this may explain why the web feature is less effective in our system. We refer to this method of using the web as *Answer Verification* to differentiate it with other approaches which attempt to answer the question on the web and then look in the target document corpus for the same answer. The latter method can result in unsupported answers. We note that the number of unsupported answers is not significantly different between runs 'ibmsqa02a' and 'ibmsqa02b' (11 vs. 8) but when we used the answer strings of confident questions as feedback to the run 'ibmsqa02c', the number of unsupported answers went up significantly (18 unsupported answers).

2.3 Statistical Machine Translation Thesaurus

Generally, an answer to a fact-seeking question can be decomposed as

$$a = a_d + a_s \qquad (2)$$

where a_d is the desired answer and a_s is the supporting evidence for the answer. Although words comprising the answer support are generally found in the question, words such as the focus of the question are sometimes deleted in the answer. Following our general approach of learning phenomena from training data, we used our question-answer corpus to train a Model 1 translation matrix (Brown et al., 1993). Questions were tokenized with casing information folded and answers were both tokenized and name entity tagged. A question answer pair is presented below before and after the prepreprocessing.

Q: How tall is Mt. Everest? *A: He started with the highest , 29,028 - foot Mt. Everest , in 1984*
Q: how tall is mt. everest ? *A: he started with the highest , 29,028 - foot mt. everest , in 1984 measure_ne*

We had 4K training pairs from the KM trivia database, 1.6K pairs from TREC8 and 10.7K pairs from TREC9. The latter were derived from correct judgements given to questions in those evaluations and which also came from

System	Description	CWS	Right	Inexact	Unsup	Wrong	Rej
ibmsqa02a	Base system	0.454	140 (28%)	37	11	312 (62.4%)	12/83
ibmsqa02b	No web or rejection	0.403	121 (24.2%)	43	8	328 (65.6%)	0
ibmsqa02c	Feedback loop	0.455	145 (29%)	44	18	293 (58.6%)	11/49

Table 1: Performance on TREC-11.

.

unique sentences in the corpus. This data was split into two and separate translation models were derived. Entries which occurred in both translation models were retained; a few of the more interesting entries are shown below in Table 2. Each word is shown with the 5 top translation candidates. For the word "who", the model prefers to see a named entity tag "person_ne" with a relative high probability. Even though the number of translation pairs is small (16.3K pairs), for the question answering application we are interested in only the most common words, which are potentially modified in the translated output of the question; rarer words have to appear identical to the form in the question. Using this additional thesaurus resource, we re-ranked the answer candidate windows (windows of text bounded by the question terms and the answer candidate) and quantized the rank into 5 bins (1,2, high, mid and low) for use in the maximum entropy answer selection module. We have not separately investigated the effect of this ranking, so details will presented in the future.

2.4 Answer Patterns

The approach described in (Soubbotin, 2001) uses patterns for locating answers. In a related work, (Ravichandran and Hovy, 2002) has shown how to extract patterns in an unsupervised manner from the web. In this work, we use the supervised corpus of question and answers to extract n-grams occurring in the answer. To specialize the pattern for a particular question type, the question was represented only by the question word and the first word to its right. To generalize the answer candidate window, it was modified to replace all non-stop question words with "<queryTerm>" and the answer candidate with "<answer>". So for the example above,

QF: how tall
MW: he started with the highest , <answer> <queryTerm> measure_ne

where *QF* stands for the question focus and *MW* stands for the mapped answer candidate window. Ideally, the question would be represented by more than just the word adjacent to the question word but in most cases this suffices. To overcome some of the limitations of this choice, we also chose features relating the predicted answer tag and an answer pattern. An answer pattern consists of 5-grams or larger chosen with a count cutoff. The total number of pattern features incorporated was 8.5K out 15.3K features.

3 Answer Selection

Answer selection was performed as we have in previous years with minor modifications. First, a fast-match technique of selecting answer sentences is used and top 100 sentences are selected. This phase yields sentences which have the answer- pattern in TREC-10 for 80% of the sentences. Considering the approximately 10% of questions which were to be rejected in TREC-10, the error of the sentence selector is about 10% with a list size of 100 sentences.

In order to select exact answers, we extracted all parse nodes which were noun phrases and together with all phrases which were named entities formed a candidate pool. As mentioned before, our system suffered a great deal of inexact answers in the judgement and these were mostly due to the decision to accept any phrase thus selected which had an answer pattern. Below we discuss some experiments in which a phrase is considered correct only if it contains only the answer pattern.

who	
a	$t(a \mid q)$
person_ne	0.125
,	0.010
the	0.051
.	0.046
”	0.042

haiti	
a	$t(a \mid q)$
haiti	0.076
port-au-prince	0.048
miami	0.034
people	0.021
haitian	0.018

river	
a	$t(a \mid q)$
river	0.217
the	0.081
water	0.060
location_ne	0.039
many	0.028

nuclear	
a	$t(a \mid q)$
nuclear	0.183
atomic	0.020
at	0.013
soviet	0.010
site	0.010

tall	
a	$t(a \mid q)$
measure_ne	0.056
foot	0.041
feet	0.027
-	0.017
i	0.012

team	
a	$t(a \mid q)$
team	0.099
organization_ne	0.056
game	0.030
;	0.029
their	0.023

Table 2: Translation entries for some question words.

.

For the training corpus of chunks with their labeled decision of correct or incorrect, we formulated features such as whether the desired named entity was found in the chunk. The features described above were added to the base model described in (Ittycheriah et al., 2001) and weights were derived using the maximum entropy algorithm. For a typical answer candidate, 50-100 features are able to fire for each decision. The answer candidate that has the highest probability is chosen for the output.

4 Rejection

For questions which are determined to have no answer in the corpus, the system was supposed to return 'NIL' as the document id. To determine which questions to reject, we employed the distribution $p(c|q,a)$ and used a threshold on the distribution. However, the system sometimes encounters events which are not sufficiently represented in the training corpus and to allow some level of control it was useful to smooth this probability with a decreasing function of chunk rank. This smooth estimate was computed as

$$p^* = (1-\alpha)p(c|q,a) + \alpha(1-0.1(\text{chunk\_rank}))$$

where chunk_rank was saturated at 10. This year the alpha was set to 0.2 and the rejection threshold to 0.3. The rejection threshold was optimized on the accuracy of TREC-10 questions using the TREC corpus of documents. We plot in Figure 1, the cumulative distribution function of questions with answers in the corpus and also 1.0 minus the cumulative distribution function for questions which should be rejected. The plot is for TREC-10 questions using the TREC corpus of documents for answers. We expected to reject about 80 answers in our base system and the actual run seems to have done approximately the same. The feedback loop of ibmsqa02c seems to have reduced the number of rejections and thus the precision of rejections has improved from 0.145 to 0.224 while maintaining the recall rate.

5 Analysis & Subsequent Experiments

One method of characterizing a test set is with respect to a set of answer tags. The primary difference between TREC-10 and TREC-11 is in the composition of the answer tags and these are presented for the top set of tags in Figure 2. The drastic difference between the test sets

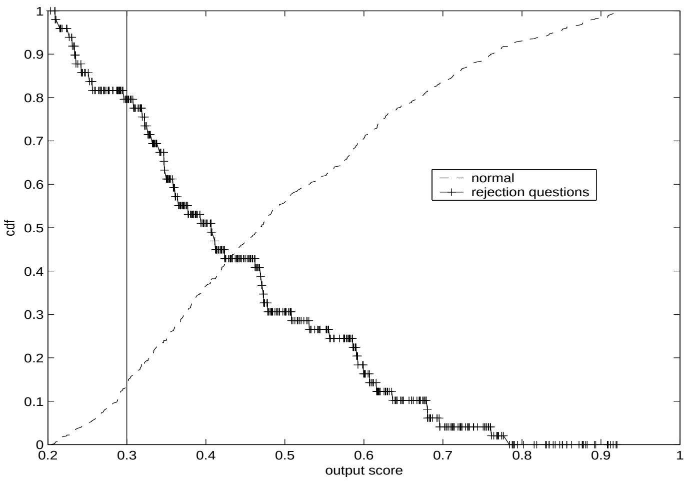

Figure 1: TREC-10 scores for normal and rejection questions.

is in the number of questions being classified PHRASE (this class represents any question which does not fall into other categories). This result reflects the reduction in definitional questions and the emphasis on exact answer questions; however, calibrating rejection rates and system strategies on TREC-10 is mismatched with the evaluation.

In order to overcome the excessive number of inexact answers produced by our system, we trained the model indicating only those phrases which exactly matched the answer pattern to be correct. As noted earlier, this is only a partial solution since some answers are now considered incorrect when they seem quite reasonable. For example in the first question of TREC-8, the answer of "Hugo Young" is now considered incorrect since the answer pattern contains only "Young". This exact match reduced the number of correct training instances about 33% (reduced from 30K to 20K where the total number of training instances is 207K); inspection of these instances indicates (a) some exact answers are now labelled incorrect, (b) majority of phrases containing the answer plus extra words are now labelled incorrect. The number of answers of type (a) is relatively small (estimated about 10% of the chunks). We then calibrated the performance of our system on TREC-11 by using the answer patterns and modifying the scoring script to accept the pattern only if

```
if ($answer_str =~ /^(\s+)?$p(\s+)?$/i)
```

The results of the system using the answer patterns are generally lower and each run seems to suffer about the same amount. Table 5 shows the results of using the new model. We emphasize that these results are obtained using the perl patterns as opposed to human judgments in the evaluation. In order to remove the effect of rejection, we modified the threshold (to 0.22 from 0.3) in the new model to output about the same number of questions rejected so that the improvement in scores is not dominated by getting only rejection questions correct. The results indicate a 46% improvement in the TREC-10 test but only about 7% gain in TREC-11. Investigating this discrepancy will be subject of future work.

In Table 3, these are answers which were accepted by the evaluation system but are now training examples for the incorrect answers. Examples of system output with the exact answer fix is shown in Table 4 with the older strings as well to demonstrate the nature of the fix. The first two examples show answers which satisfy the answer patterns exactly at test time. The last two example show errors by the system, but

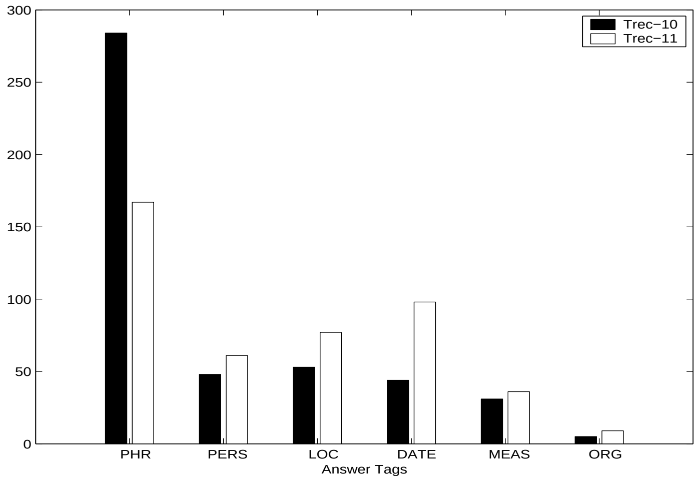

Figure 2: Comparison of answer tags between TREC-10 and 11.

What canine was made famous by Eric Knight?	Lassie Come - Home
Professor Moriarty was whose rival?	Sherlock Holmes' nemesis
What is Francis Scott Key best known for?	write the Star Spangled Banner

Table 3: Training data instances which are rejected for the exact answer fix.

.

Qnum	Question	Old Answer	Answer
1059	What peninsula is Spain part of?	position on the Iberian Peninsula	Iberian Peninsula
1215	When was President Kennedy shot?	shot on Nov. 22 , 1963	Nov. 22 , 1963
1316	What was the name of the plane Lindbergh flew solo across the Atlantic?	Spirit of St. Louis	Charles A.
1348	How cold should a refrigerator be?	28 degrees Farenheit	soda ice cold

Table 4: TREC-10 QA pairs before and after the exact answer fix.

.

System	Description	CWS	Right	Wrong	Rej
trec10-perl	Base system	0.289	92	408	10/76
trec10-perl	Exact answer fix	0.423	127	373	16/92
ibmsqa02a-perl	Base system	0.438	134	366	12/83
ibmsqa02a-perl	Exact answer fix	0.469	144	356	4/52

Table 5: Experimental results using perl-patterns since TREC-11 evaluation.

.

overall the system was able to produce more answers which satisfied the exact match criteria.

6 Conclusions and Future Work

In TREC-11, our method of selecting which candidates were exact answers did not satisfy the exact match criteria of the evaluation. We have since modified our system to extract exact answers and retrained the system. We incorporated two novel concepts (a statistical machine translation thesaurus and lexical patterns derived from supervised question-answer pairs) since last year.

In TREC-11, although we thresholded the distribution $p(c|q, a)$ to reject answers, this we recognize as being deficient in the following sense. We should recognize a question as not having an answer in the corpus by taking into consideration all the answers found and not just the top ranking answer.

7 Acknowledgement

This work is supported in part by DARPA under SPAWAR contract number N66001-99-2-8916.

References

Adam L. Berger, Vincent Della Pietra, and Stephen Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

E. Brill, J. Lin, M. Banko, S. Dumais, and A. Ng. 2001. Data-intensive question answering. *TREC-10 Proceedings*, pages 393–400.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Jennifer Chu-Carroll, John Prager, Christopher Welty, Krzysztof Czuba, and David Ferruci. 2002. A multi-strategy and multi-source approach to question answering. *To appear in TREC-11 Proceedings.*

C.L.A. Clarke, G.V. Cormack, T.R. Lynam, C.M. Li, and G.L. McLearn. 2001. Web reinforced question answering (multitext experiments for trec 2001). *TREC-10 Proceedings*, pages 673–679.

Sanda Harabagiu and et. al. 2000. Falcon: Boosting knowledge for answer engines. *TREC-9 Proceedings*, pages 50–59.

Abraham Ittycheriah, Martin Franz, and Salim Roukos. 2001. IBM's statistical question answering system – trec-10. *TREC-10 Proceedings*, pages 258–264.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 41–47.

M. M. Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. *TREC-10 Proceedings*, pages 293–302.

Rohini Srihari and Wei Li. 1999. Question answering supported by information extraction. *TREC-8 Proceedings*, pages 75–85.

Ellen M. Voorhees and Dawn M. Tice. 1999. The TREC-8 question answering track evaluation. *TREC-8 Proceedings*, pages 41–63, Nov.