Question Answering as Question-Biased Term Extraction: A New Approach toward Multilingual QA

Yutaka Sasaki
Department of Natural Language Processing
ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288 Japan
yutaka.sasaki@atr.jp

Abstract

This paper regards *Question Answering (QA)* as *Question-Biased Term Extraction (QBTE)*. This new QBTE approach liberates QA systems from the heavy burden imposed by *question types* (or *answer types*). In conventional approaches, a QA system analyzes a given question and determines the question type, and then it selects answers from among answer candidates that match the question type. Consequently, the output of a QA system is restricted by the design of the question types. The QBTE directly extracts answers as terms biased by the question. To confirm the feasibility of our QBTE approach, we conducted experiments on the CRL QA Data based on 10-fold cross validation, using *Maximum Entropy Models (MEMs)* as an ML technique. Experimental results showed that the trained system achieved 0.36 in MRR and 0.47 in Top5 accuracy.

1 Introduction

The conventional *Question Answering (QA)* architecture is a cascade of the following building blocks:

Question Analyzer analyzes a question sentence and identifies the *question types* (or *answer types*).

Document Retriever retrieves documents related to the question from a large-scale document set.

Answer Candidate Extractor extracts answer candidates that match the question types from the retrieved documents.

Answer Selector ranks the answer candidates according to the syntactic and semantic conformity of each answer with the question and its context in the document.

Typically, question types consist of *named entities*, e.g., PERSON, DATE, and ORGANIZATION, *numerical expressions*, e.g., LENGTH, WEIGHT, SPEED, and *class names*, e.g., FLOWER, BIRD, and FOOD. The question type is also used for selecting answer candidates. For example, if the question type of a given question is PERSON, the answer candidate extractor lists only person names that are tagged as the named entity PERSON.

The conventional QA architecture has a drawback in that the question-type system restricts the range of questions that can be answered by the system. It is thus problematic for QA system developers to carefully design and build an answer candidate extractor that works well in conjunction with the question-type system. This problem is particularly difficult when the task is to develop a multilingual QA system to handle languages that are unfamiliar to the developer. Developing high-quality tools that can extract named entities, numerical expressions, and class names for each foreign language is very costly and time-consuming.

Recently, some pioneering studies have investigated approaches to automatically construct QA components from scratch by applying machine learning techniques to training data (Ittycheriah et al., 2001a)(Ittycheriah et al., 2001b)(Ng et al., 2001) (Pasca and Harabagiu)(Suzuki et al., 2002)(Suzuki

Proceedings of the 43rd Annual Meeting of the ACL, pages 215–222,

Table 1: Number of Questions in Question Types of CRL QA Data

# of Questions	# of Question Types	Example
1-9	74	AWARD, CRIME, OFFENSE
10-50	32	PERCENT, N_PRODUCT, YEAR_PERIOD
51-100	6	COUNTRY, COMPANY, GROUP
100-300	3	PERSON, DATE, MONEY
Total	115	

et al., 2003) (Zukerman and Horvitz, 2001)(Sasaki et al., 2004). These approaches still suffer from the problem of preparing an adequate amount of training data specifically designed for a particular QA system because each QA system uses its own question-type system. It is very typical in the course of system development to redesign the question-type system in order to improve system performance. This inevitably leads to revision of a large-scale training dataset, which requires a heavy workload.

For example, assume that you have to develop a Chinese or Greek QA system and have 10,000 pairs of question and answers. You have to manually classify the questions according to your own question-type system. In addition, you have to annotate the tags of the question types to large-scale Chinese or Greek documents. If you wanted to redesign the question type ORGANIZATION to three categories, COMPANY, SCHOOL, and OTHER_ORGANIZATION, then the ORGANIZATION tags in the annotated document set would need to be manually revisited and revised.

To solve this problem, this paper regards Question Answering as *Question-Biased Term Extraction (QBTE)*. This new QBTE approach liberates QA systems from the heavy burden imposed by question types.

Since it is a challenging as well as a very complex and sensitive problem to directly extract answers without using question types and only using features of questions, correct answers, and contexts in documents, we have to investigate the feasibility of this approach: how well can answer candidates be extracted, and how well are answer candidates ranked?

In response, this paper employs the machine learning technique *Maximum Entropy Models (MEMs)* to extract answers to a question from documents based on question features, document features, and the combined features. Experimental results show the performance of a QA system that applies MEMs.

2 Preparation

2.1 Training Data

Document Set Japanese newspaper articles of The Mainichi Newspaper published in 1995.

Question/Answer Set We used the CRL[1] QA Data (Sekine et al., 2002). This dataset comprises 2,000 Japanese questions with correct answers as well as question types and IDs of articles that contain the answers. Each question is categorized as one of 115 hierarchically classified question types.

The document set is used not only in the training phase but also in the execution phrase.

Although the CRL QA Data contains question types, *the information of question types are not used for the training*. This is because more than the 60% of question types have fewer than 10 questions as examples (Table 1). This means it is very unlikely that we can train a QA system that can handle this 60% due to data sparseness. [2] Only for the purpose of analyzing experimental results in this paper do we refer to the question types of the dataset.

2.2 Learning with Maximum Entropy Models

This section briefly introduces the machine learning technique Maximum Entropy Models and describes how to apply MEMs to QA tasks.

2.2.1 Maximum Entropy Models

Let $\mathcal{X}$ be a set of input symbols and $\mathcal{Y}$ be a set of class labels. A sample (x, y) is a pair of input x=$\{x_1,\dots,x_m\}$ ($x_i \in \mathcal{X}$) and output $y \in \mathcal{Y}$.

[1] Presently, National Institute of Information and Communications Technology (NICT), Japan

[2] A machine learning approach to hierarchical question analysis was reported in (Suzuki et al., 2003), but training and maintaining an answer extractor for question types of fine granularity is not an easy task.

The Maximum Entropy Principle (Berger et al., 1996) is to find a model $p* = \underset{p \in C}{\operatorname{argmax}} H(p)$, which means a probability model $p(y|x)$ that maximizes entropy $H(p)$.

Given data $(x^{(1)}, y^{(1)})$,...,$(x^{(n)}, y^{(n)})$, let $\bigcup_k (x^{(k)} \times \{y^{(k)}\}) = \{\langle \tilde{x}_1, \tilde{y}_1 \rangle, ..., \langle \tilde{x}_i, \tilde{y}_i \rangle, ..., \langle \tilde{x}_m, \tilde{y}_m \rangle\}$. This means that we enumerate all pairs of an input symbol and label and represent them as $\langle \tilde{x}_i, \tilde{y}_i \rangle$ using index i $(1 \leq i \leq m)$.

In this paper, feature function f_i is defined as follows.

$$f_i(x, y) = \begin{cases} 1 & if\ \ \tilde{x}_i \in x\ \ and\ \ y = \tilde{y}_i \\ 0 & otherwise \end{cases}$$

We use all combinations of input symbols in x and class labels for features (or the feature function) of MEMs.

With Lagrangian $\lambda = \lambda_1, ..., \lambda_m$, the dual function of H is:

$$\Psi(\lambda) = -\sum_x \tilde{p}(x) \log Z_\lambda(x) + \sum \lambda_i \tilde{p}(f_i),$$

where $Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y))$ and $\tilde{p}(x)$ and $\tilde{p}(f_i)$ indicate the empirical distribution of x and f_i in the training data.

The dual optimization problem $\lambda* = \underset{\lambda}{\operatorname{argmax}} \Psi(\lambda)$ can be efficiently solved as an optimization problem without constraints. As a result, probabilistic model $p* = p_\lambda*$ is obtained as:

$$p_{\lambda*}(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x, y)\right).$$

2.2.2 Applying MEMs to QA

Question analysis is a classification problem that classifies questions into different question types. Answer candidate extraction is also a classification problem that classifies words into answer types (i.e., question types), such as PERSON, DATE, and AWARD. Answer selection is an exactly classification that classifies answer candidates as positive or negative. Therefore, we can apply machine learning techniques to generate classifiers that work as components of a QA system.

In the QBTE approach, these three components, i.e., question analysis, answer candidate extraction, and answer selection, are integrated into one classifier.

To successfully carry out this goal, we have to extract features that reflect properties of correct answers of a question in the context of articles.

3 QBTE Model 1

This section presents a framework, QBTE Model 1, to construct a QA system from question-answer pairs based on the QBTE Approach. When a user gives a question, the framework finds answers to the question in the following two steps.

Document Retrieval retrieves the top N articles or paragraphs from a large-scale corpus.

QBTE creates input data by combining the question features and documents features, evaluates the input data, and outputs the top M answers.[3]

Since this paper focuses on QBTE, this paper uses a simple *idf* method in document retrieval.

Let w_i be words and w_1,w_2,...w_m be a document. Question Answering in the QBTE Model 1 involves directly classifying words w_i in the document into answer words or non-answer words. That is, given input $x^{(i)}$ for w_i, its class label is selected from among $\{I, O, B\}$ as follows:

I: if the word is in the middle of the answer word sequence;

O: if the word is not in the answer word sequence;

B: if the word is the start word of the answer word sequence.

The class labeling system in our experiment is IOB2 (Sang, 2000), which is a variation of IOB (Ramshaw and Marcus, 1995).

Input $x^{(i)}$ of each word is defined as described below.

3.1 Feature Extraction

This paper employs three groups of features as features of input data:

- Question Feature Set (QF);
- Document Feature Set (DF);
- Combined Feature Set (CF), i.e., combinations of question and document features.

[3] In this paper, M is set to 5.

3.1.1 Question Feature Set (QF)

A Question Feature Set (QF) is a set of features extracted only from a question sentence. This feature set is defined as belonging to a question sentence.

The following are elements of a Question Feature Set:

qw: an enumeration of the word n-grams ($1 \leq n \leq N$), e.g., given question "What is CNN?", the features are $\{$*qw:What, qw:is, qw:CNN, qw:What-is, qw:is-CNN* $\}$ if $N = 2$,

qq: interrogative words (e.g., who, where, what, how many),

qm1: POS1 of words in the question, e.g., given "What is CNN?", $\{$ *qm1:wh-adv, qm1:verb, qm1:noun* $\}$ are features,

qm2: POS2 of words in the question,

qm3: POS3 of words in the question,

qm4: POS4 of words in the question.

POS1-POS4 indicate part-of-speech (POS) of the IPA POS tag set generated by the Japanese morphological analyzer ChaSen. For example, "Tokyo" is analyzed as POS1 = *noun*, POS2 = *propernoun*, POS3 = *location*, and POS4 = *general*. This paper used up to 4-grams for qw.

3.1.2 Document Feature Set (DF)

Document Feature Set (DF) is a feature set extracted only from a document. Using only DF corresponds to *unbiased* Term Extraction (TE).

For each word w_i, the following features are extracted:

dw–k,. . .,dw+0,. . .,dw+k: k preceding and following words of the word w_i, e.g., $\{$ *dw–1*:w_{i-1}, *dw+0*:w_i, *dw+1*:$w_{i+1}\}$ if $k = 1$,

dm1–k,. . .,dm1+0,. . .,dm1+k: POS1 of k preceding and following words of the word w_i,

dm2–k,. . .,dm2+0,. . .,dm2+k: POS2 of k preceding and following words of the word w_i,

dm3–k,. . .,dm3+0,. . .,dm3+k: POS3 of k preceding and following words of the word w_i,

dm4–k,. . .,dm4+0,. . .,dm4+k: POS4 of k preceding and following words of the word w_i.

In this paper, k is set to 3 so that the window size is 7.

3.1.3 Combined Feature Set (CF)

Combined Feature Set (CF) contains features created by combining question features and document features. QBTE Model 1 employs CF. For each word w_i, the following features are created.

cw–k,. . .,cw+0,. . .,cw+k: matching results (true/false) between each of dw–k,...,dw+k features and any qw feature, e.g., *cw–1:true* if *dw–1:President* and *qw: President*,

cm1–k,. . .,cm1+0,. . .,cm1+k: matching results (true/false) between each of dm1–k,...,dm1+k features and any POS1 in qm1 features,

cm2–k,. . .,cm2+0,. . .,cm2+k: matching results (true/false) between each of dm2–k,...,dm2+k features and any POS2 in qm2 features,

cm3–k,. . .,cm3+0,. . .,cm3+k: matching results (true/false) between each of dm3–k,...,dm3+k features and any POS3 in qm3 features,

cm4–k,. . .,cm4+0,. . .,cm4+k: matching results (true/false) between each of dm4–k,...,dm4+k features and any POS4 in qm4 features,

cq–k,. . .,cq+0,. . .,cq+k: combinations of each of dw–k,...,dw+k features and qw features, e.g., *cq–1:President&Who* is a combination of *dw–1:President* and *qw:Who*.

3.2 Training and Execution

The training phase estimates a probabilistic model from training data ($x^{(1)}$,$y^{(1)}$),...,($x^{(n)}$,$y^{(n)}$) generated from the CRL QA Data. The execution phase evaluates the probability of $y'^{(i)}$ given input$x'^{(i)}$ using the the probabilistic model.

Training Phase

1. Given question q, correct answer a, and document d.
2. Annotate $\langle A\rangle$ and $\langle /A\rangle$ right before and after answer a in d.
3. Morphologically analyze d.
4. For $d = w_1, ..., \langle A\rangle, w_j, ..., w_k, \langle /A\rangle, ..., w_m$, extract features as $x^{(1)}$,...,$x^{(m)}$.
5. Class label $y^{(i)} = B$ if w_i follows $\langle A\rangle$, $y^{(i)} = I$ if w_i is inside of $\langle A\rangle$ and $\langle /A\rangle$, and $y^{(i)} = O$ otherwise.

Table 2: Main Results with 10-fold Cross Validation

	Correct Answer Rank					MRR	Top5
	1	2	3	4	5		
Exact match	453	139	68	35	19	0.28	0.36
Partial match	684	222	126	80	48	0.43	0.58
Ave.						0.355	0.47
Manual evaluation	578	188	86	55	34	0.36	0.47

6. Estimate $p_\lambda *$ from $(x^{(1)},y^{(1)}),...,(x^{(n)},y^{(n)})$ using Maximum Entropy Models.

The execution phase extracts answers from retrieved documents as Term Extraction, biased by the question.

Execution Phase

1. Given question q and paragraph d.
2. Morphologically analyze d.
3. For w_i of $d = w_1, ..., w_m$, create input data $x'^{(i)}$ by extracting features.
4. For each $y'^{(j)} \in \mathcal{Y}$, compute $p_\lambda * (y'^{(j)}|x'^{(i)})$, which is a probability of $y'^{(j)}$ given $x'^{(i)}$.
5. For each $x'^{(i)}$, $y'^{(j)}$ with the highest probability is selected as the label of w_i.
6. Extract word sequences that start with the word labeled B and are followed by words labeled I from the labeled word sequence of d.
7. Rank the top M answers according to the probability of the first word.

This approach is designed to extract only the most highly probable answers. However, pin-pointing only answers is not an easy task. To select the top five answers, it is necessary to loosen the condition for extracting answers. Therefore, in the execution phase, we only give label O to a word if its probability exceeds 99%, otherwise we give the second most probable label.

As a further relaxation, word sequences that include B inside the sequences are extracted for answers. This is because our preliminary experiments indicated that it is very rare for two answer candidates to be adjacent in Question-Biased Term Extraction, unlike an ordinary Term Extraction task.

4 Experimental Results

We conducted 10-fold cross validation using the CRL QA Data. The output is evaluated using the Top5 score and MRR.

Top5 Score shows the rate at which at least one correct answer is included in the top 5 answers.

MRR (Mean Reciprocal Rank) is the average reciprocal rank $(1/n)$ of the highest rank n of a correct answer for each question.

Judgment of whether an answer is correct is done by both automatic and manual evaluation. Automatic evaluation consists of exact matching and partial matching. Partial matching is useful for absorbing the variation in extraction range. A partial match is judged correct if a system's answer completely includes the correct answer or the correct answer completely includes a system's answer. Table 2 presents the experimental results. The results show that a QA system can be built by using our QBTE approach. The manually evaluated performance scored MRR=0.36 and Top5=0.47. However, manual evaluation is costly and time-consuming, so we use automatic evaluation results, i.e., exact matching results and partial matching results, as a pseudo lower-bound and upper-bound of the performances. Interestingly, the manual evaluation results of MRR and Top5 are nearly equal to the average between exact and partial evaluation.

To confirm that the QBTE ranks potential answers to the higher rank, we changed the number of paragraphs retrieved from a large corpus from $N = 1, 3, 5$ to 10. Table 3 shows the results. Whereas the performances of Term Extraction (TE) and Term Extraction with question features (TE+QF) significantly degraded, the performance of the QBTE (CF) did not severely degrade with the larger number of retrieved paragraphs.

Table 3: Answer Extraction from Top N documents

Feature set	Top N paragraphs	Match	Correct Answer Rank					MRR	Top5
			1	2	3	4	5		
TE (DF)	1	Exact	102	109	80	71	62	**0.11**	**0.21**
		Partial	207	186	155	153	121	**0.21**	**0.41**
	3	Exact	65	63	55	53	43	0.07	0.14
		Partial	120	131	112	108	94	0.13	0.28
	5	Exact	51	38	38	36	36	0.05	0.10
		Partial	99	80	89	81	75	0.10	0.21
	10	Exact	29	17	19	22	18	0.03	0.07
		Partial	59	38	35	49	46	0.07	0.14
TE (DF) + QF	1	Exact	120	105	94	63	80	**0.12**	**0. 23**
		Partial	207	198	175	126	140	**0.21**	**0 .42**
	3	Exact	65	68	52	58	57	0.07	0.15
		Partial	119	117	111	122	106	0.13	0.29
	5	Exact	44	57	41	35	31	0.05	0.10
		Partial	91	104	71	82	63	0.10	0.21
	10	Exact	28	42	30	28	26	0.04	0.08
		Partial	57	68	57	56	45	0.07	0.14
QBTE (CF)	1	Exact	453	139	68	35	19	**0.28**	0.36
		Partial	684	222	126	80	48	**0.43**	0.58
	3	Exact	403	156	92	52	43	0.27	**0.37**
		Partial	539	296	145	105	92	0.42	**0.62**
	5	Exact	381	153	92	59	50	0.26	**0.37**
		Partial	542	291	164	122	102	0.40	0.61
	10	Exact	348	128	92	65	57	0.24	0.35
		Partial	481	257	173	124	102	0.36	0.57

5 Discussion

Our approach needs no question type system, and it still achieved 0.36 in MRR and 0.47 in Top5. This performance is comparable to the results of SAIQA-II (Sasaki et al., 2004) (MRR=0.4, Top5=0.55) whose question analysis, answer candidate extraction, and answer selection modules were independently built from a QA dataset and an NE dataset, which is limited to eight named entities, such as PERSON and LOCATION. Since the QA dataset is not publicly available, it is not possible to directly compare the experimental results; however we believe that the performance of the QBTE Model 1 is comparable to that of the conventional approaches, even though it does not depend on question types, named entities, or class names.

Most of the partial answers were judged correct in manual evaluation. For example, for “How many times bigger ...?”, “two times” is a correct answer but “two” was judged correct. Suppose that “John Kerry” is a prepared correct answer in the CRL QA Data. In this case, “Senator John Kerry” would also be correct. Such additions and omissions occur because our approach is not restricted to particular extraction units, such as named entities or class names.

The performance of QBTE was affected little by the larger number of retrieved paragraphs, whereas the performances of TE and TE + QF significantly degraded. This indicates that QBTE Model 1 is not mere Term Extraction with document retrieval but Term Extraction appropriately biased by questions.

Our experiments used no information about question types given in the CRL QA Data because we are seeking a universal method that can be used for any QA dataset. *Beyond this main goal*, as a reference, The Appendix shows our experimental results classified into question types without using them in the training phase. The results of automatic evaluation of complete matching are in Top5 (T5), and MRR and partial matching are in Top5 (T5’) and MRR’. It is interesting that minor question types were correctly answered, e.g., SEA and WEAPON, for which there was only one training question.

We also conducted an additional experiment, as a reference, on the training data that included question types defined in the CRL QA Data; the question-type of each question is added to the qw feature. The performance of QBTE from the first-ranked paragraph showed no difference from that of experiments shown in Table 2.

6 Related Work

There are two previous studies on integrating QA components into one using machine learning/statistical NLP techniques. Echihabi et al. (Echihabi et al., 2003) used Noisy-Channel Models to construct a QA system. In this approach, the range of Term Extraction is not trained by a data set but selected from answer candidates, e.g., named entities and noun phrases, generated by a decoder. Lita et al. (Lita and Carbonell, 2004) share our motivation to build a QA system only from question-answer pairs without depending on the question types. Their method finds clusters of questions and defines how to answer questions in each cluster. However, their approach is to find snippets, i.e., short passages including answers, not exact answers extracted by Term Extraction.

7 Conclusion

This paper described a novel approach to extracting answers to a question using probabilistic models constructed from only question-answer pairs. This approach requires no question type system, no named entity extractor, and no class name extractor. To the best of our knowledge, no previous study has regarded Question Answering as Question-Biased Term Extraction. As a feasibility study, we built a QA system using Maximum Entropy Models on a 2000-question/answer dataset. The results were evaluated by 10-fold cross validation, which showed that the performance is 0.36 in MRR and 0.47 in Top5. Since this approach relies on a morphological analyzer, applying the QBTE Model 1 to QA tasks of other languages is our future work.

Acknowledgment

This research was supported by a contract with the National Institute of Information and Communications Technology (NICT) of Japan entitled, “A study of speech dialogue translation technology based on a large corpus”.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra: A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, Vol. 22, No. 1, pp. 39–71 (1996).

Abdessamad Echihabi and Daniel Marcu: A Noisy-Channel Approach to Question Answering, *Proc. of ACL-2003*, pp. 16-23 (2003).

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi: Question Answering Using Maximum-Entropy Components, *Proc. of NAACL-2001* (2001).

Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, and Adwait Ratnaparkhi: IBM’s Statistical Question Answering System – TREC-10, *Proc. of TREC-10* (2001).

Lucian Vlad Lita and Jaime Carbonell: Instance-Based Question Answering: A Data-Driven Approach: *Proc. of EMNLP-2004*, pp. 396–403 (2004).

Hwee T. Ng, Jennifer L. P. Kwan, and Yiyuan Xia: Question Answering Using a Large Text Database: A Machine Learning Approach: *Proc. of EMNLP-2001*, pp. 67–73 (2001).

Marisu A. Pasca and Sanda M. Harabagiu: High Performance Question/Answering, *Proc. of SIGIR-2001*, pp. 366–374 (2001).

Lance A. Ramshaw and Mitchell P. Marcus: Text Chunking using Transformation-Based Learning, *Proc. of WVLC-95*, pp. 82–94 (1995).

Erik F. Tjong Kim Sang: Noun Phrase Recognition by System Combination, *Proc. of NAACL-2000*, pp. 55–55 (2000).

Yutaka Sasaki, Hideki Isozaki, Jun Suzuki, Kouji Kokuryou, Tsutomu Hirao, Hideto Kazawa, and Eisaku Maeda, SAIQA-II: A Trainable Japanese QA System with SVM, *IPSJ Journal*, Vol. 45, NO. 2, pp. 635-646, 2004. (in Japanese)

Satoshi Sekine, Kiyoshi Sudo, Yusuke Shinyama, Chikashi Nobata, Kiyotaka Uchimoto, and Hitoshi Isahara, NYU/CRL QA system, QAC question analysis and CRL QA data, *in Working Notes of NTCIR Workshop 3* (2002).

Jun Suzuki, Yutaka Sasaki, and Eisaku Maeda: SVM Answer Selection for Open-Domain Question Answering, *Proc. of Coling-2002*, pp. 974–980 (2002).

Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda: Directed Acyclic Graph Kernel, *Proc. of ACL 2003 Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, pp. 61–68, Sapporo (2003).

Ingrid Zukerman and Eric Horvitz: Using Machine Learning Techniques to Interpret WH-Questions, *Proc. of ACL-2001*, Toulouse, France, pp. 547–554 (2001).

Appendix: Analysis of Evaluation Results w.r.t. Question Type — Results of QBTE from the first-ranked paragraph (NB: No information about these question types was used in the training phrase.)

Question Type	#Qs	MRR	T5	MRR'	T5'
GOE	36	0.30	0.36	0.41	0.53
GPE	4	0.50	0.50	1.00	1.00
N_EVENT	7	0.76	0.86	0.76	0.86
EVENT	19	0.17	0.21	0.41	0.53
GROUP	74	0.28	0.35	0.45	0.62
SPORTS_TEAM	15	0.28	0.40	0.45	0.73
BROADCAST	1	0.00	0.00	0.00	0.00
POINT	2	0.00	0.00	0.00	0.00
DRUG	2	0.00	0.00	0.00	0.00
SPACESHIP	4	0.88	1.00	0.88	1.00
ACTION	18	0.22	0.22	0.30	0.44
MOVIE	6	0.50	0.50	0.56	0.67
MUSIC	8	0.19	0.25	0.56	0.62
WATER_FORM	3	0.50	0.67	0.50	0.67
CONFERENCE	17	0.14	0.24	0.46	0.65
SEA	1	1.00	1.00	1.00	1.00
PICTURE	1	0.00	0.00	0.00	0.00
SCHOOL	21	0.10	0.10	0.33	0.43
ACADEMIC	5	0.20	0.20	0.37	0.60
PERCENT	47	0.35	0.43	0.43	0.55
COMPANY	77	0.45	0.55	0.57	0.70
PERIODX	1	1.00	1.00	1.00	1.00
RULE	35	0.30	0.43	0.49	0.69
MONUMENT	2	0.00	0.00	0.25	0.50
SPORTS	9	0.17	0.22	0.40	0.67
INSTITUTE	26	0.38	0.46	0.53	0.69
MONEY	110	0.33	0.40	0.48	0.63
AIRPORT	4	0.38	0.50	0.44	0.75
MILITARY	4	0.00	0.00	0.25	0.25
ART	4	0.25	0.50	0.25	0.50
MONTH_PERIOD	6	0.06	0.17	0.06	0.17
LANGUAGE	3	1.00	1.00	1.00	1.00
COUNTX	10	0.33	0.40	0.38	0.60
AMUSEMENT	2	0.00	0.00	0.00	0.00
PARK	1	0.00	0.00	0.00	0.00
SHOW	3	0.78	1.00	1.11	1.33
PUBLIC_INST	19	0.18	0.26	0.34	0.53
PORT	3	0.17	0.33	0.33	0.67
N_COUNTRY	8	0.28	0.38	0.32	0.50
NATIONALITY	4	0.50	0.50	1.00	1.00
COUNTRY	84	0.45	0.60	0.51	0.67
OFFENSE	9	0.23	0.44	0.23	0.44
CITY	72	0.41	0.50	0.53	0.65
N_FACILITY	4	0.25	0.25	0.38	0.50
FACILITY	11	0.20	0.36	0.25	0.55
TIMEX	3	0.00	0.00	0.00	0.00
TIME_TOP	2	0.00	0.00	0.50	0.50
TIME_PERIOD	8	0.12	0.12	0.48	0.75
TIME	13	0.22	0.31	0.29	0.38
ERA	3	0.00	0.00	0.33	0.33
PHENOMENA	5	0.50	0.60	0.60	0.80
DISASTER	4	0.50	0.75	0.50	0.75
OBJECT	5	0.47	0.60	0.47	0.60
CAR	1	1.00	1.00	1.00	1.00
RELIGION	5	0.30	0.40	0.30	0.40
WEEK_PERIOD	4	0.05	0.25	0.55	0.75
WEIGHT	12	0.21	0.25	0.31	0.42
PRINTING	6	0.17	0.17	0.38	0.50

Question Type	#Q	MRR	T5	MRR'	T5'
RANK	7	0.18	0.29	0.54	0.71
BOOK	6	0.31	0.50	0.47	0.67
AWARD	9	0.17	0.33	0.34	0.56
N_LOCATION	2	0.10	0.50	0.10	0.50
VEGETABLE	10	0.31	0.50	0.34	0.60
COLOR	5	0.20	0.20	0.20	0.20
NEWSPAPER	7	0.61	0.71	0.61	0.71
WORSHIP	8	0.47	0.62	0.62	0.88
SEISMIC	1	0.00	0.00	1.00	1.00
N_PERSON	72	0.30	0.39	0.43	0.60
PERSON	282	0.18	0.21	0.46	0.55
NUMEX	19	0.32	0.32	0.35	0.47
MEASUREMENT	1	0.00	0.00	0.00	0.00
P_ORGANIZATION	3	0.33	0.33	0.67	0.67
P_PARTY	37	0.30	0.41	0.43	0.57
GOVERNMENT	37	0.50	0.54	0.53	0.57
N_PRODUCT	41	0.25	0.37	0.37	0.56
PRODUCT	58	0.24	0.34	0.44	0.69
WAR	2	0.75	1.00	0.75	1.00
SHIP	7	0.26	0.43	0.40	0.57
N_ORGANIZATION	20	0.14	0.25	0.28	0.55
ORGANIZATION	23	0.08	0.13	0.20	0.30
SPEED	1	0.00	0.00	1.00	1.00
VOLUME	5	0.00	0.00	0.18	0.60
GAMES	8	0.28	0.38	0.34	0.50
POSITION_TITLE	39	0.20	0.28	0.30	0.44
REGION	22	0.17	0.23	0.46	0.64
GEOLOGICAL	3	0.42	0.67	0.42	0.67
LOCATION	2	0.00	0.00	0.50	0.50
EXTENT	22	0.04	0.09	0.13	0.18
CURRENCY	1	0.00	0.00	0.00	0.00
STATION	3	0.50	0.67	0.50	0.67
RAILROAD	1	0.00	0.00	0.25	1.00
PHONE	1	0.00	0.00	0.00	0.00
PROVINCE	36	0.30	0.33	0.45	0.50
N_ANIMAL	3	0.11	0.33	0.22	0.67
ANIMAL	10	0.26	0.50	0.31	0.60
ROAD	1	0.00	0.00	0.50	1.00
DATE_PERIOD	9	0.11	0.11	0.33	0.33
DATE	130	0.24	0.32	0.41	0.58
YEAR_PERIOD	34	0.22	0.29	0.38	0.59
AGE	22	0.34	0.45	0.44	0.59
MULTIPLICATION	9	0.39	0.44	0.56	0.67
CRIME	4	0.75	0.75	0.75	0.75
AIRCRAFT	2	0.00	0.00	0.25	0.50
MUSEUM	3	0.33	0.33	0.33	0.33
DISEASE	18	0.29	0.50	0.43	0.72
FREQUENCY	13	0.18	0.31	0.19	0.38
WEAPON	1	1.00	1.00	1.00	1.00
MINERAL	18	0.16	0.22	0.25	0.39
METHOD	29	0.39	0.48	0.48	0.62
ETHNIC	3	0.42	0.67	0.75	1.00
NAME	5	0.20	0.20	0.40	0.40
SPACE	4	0.50	0.50	0.50	0.50
THEORY	1	0.00	0.00	0.00	0.00
LANDFORM	5	0.13	0.40	0.13	0.40
TRAIN	2	0.17	0.50	0.17	0.50
	2000	0.28	0.36	0.43	0.58