Answer filtering via Text Categorization in Question Answering Systems

Alessandro Moschitti
University of Texas at Dallas
Human Language Technology Research Institute
Richardson, TX 75083-0688, USA
alessandro.moschitti@utdallas.edu

Abstract

Modern Information Technologies and Web-based services are faced with the problem of selecting, filtering and managing growing amounts of textual information to which access is usually critical. On one hand, Text Categorization models allow users to browse more easily the set of texts of their own interests, by navigating in category hierarchies. On the other hand, Question/Answering is a method of retrieving information from vast document collections. In spite of their shared goal, these two information retrieval techniques have been ever applied separately.

In this paper we present a Question/Answering system that takes advantage from category information by exploiting several models of question and answer categorization. Knowing the question category has the potential of enhancing a more efficient answer extraction mechanism as the matching of the question category with the answer category allows to (1) re-rank the answers; and (2) eliminate incorrect answers. Experimental results show the effects of question and answer categorization on the overall Question Answering performance.

1. Introduction

One method of retrieving information from vast document collections is by using textual *Question/Answering*. Q/A is an Information Retrieval (IR) paradigm that returns a short list of answers, extracted from relevant documents, to a question formulated in natural language. Another, different method to find the desired information is by navigating along subject categories assigned hierarchically to groups of documents, in a style made popular by *Yahoo.com* among others. When the defined category is reached, documents are inspected and the information is eventually retrieved.

Q/A systems incorporate a paragraph retrieval engine, to find paragraphs that contain candidate answers, as reported in [3, 10]. To our knowledge no information on the text categories of these paragraphs is currently employed in any of the Q/A systems. Instead, another semantic information, such as the semantic classes of the expected answers, derived from the question processing, is used to retrieve paragraphs and later to extract answers. Typically, the semantic classes of answers are organized in hierarchical ontologies and do not relate in any way to the categories associated with documents.

The ontology of expected answer classes contains concepts like PERSON, LOCATION or PRODUCT, whereas categories associated with documents are more similar to topics than concepts, e.g., *acquisitions*, *trading* or *earnings*. Given that text categories indicate a different semantic information than the classes of the expected answers, we argue in this paper that text categories can be used to improve the quality of textual Q/A. In fact, by assigning text categories to both questions and answers, we have an additional information on their similarity. This allows us to filter out many incorrect answers and to improve the ranking of answers produced by Q/A systems.

In order to correctly assign categories to questions and answers the set of documents, on which the Q/A systems operate, has to be pre-categorized. For our experiments we trained our basic Q/A system on the well known text categorization benchmark, the *Reuters-21578*. This allows us to assume as categories of an answer the categories of the documents which contain such answer. More difficult, instead, is the assigning of categories to questions as: (a) they are not known in advance and (b) their reduced size (in term of number of words) often prevents the detection of the their categories.

The article is organized as follows: Section 2 describes our Q/A system whereas Section 3 shows the question categorization problem and the solutions adopted. Section 4 presents the filtering and the re-ranking methods that combines the basic Q/A with the question classification models. Section 5 reports the experiments on question categorization, basic Question Answering and Question Answering based on Text Categorization (TC). Finally, Section 6 derives the conclusions.

2. Textual Question Answering

The typical architecture of a Q/A system is illustrated in Figure 1. Given a question, it is first processed to determine (a) the semantic class of the expected answer and (b) what keywords constitute the queries used to retrieve relevant paragraphs. Question processing relies on external resources to identify the class of the expected answer, typically in the form of semantic ontologies (*Answer Type Ontology*). The semantic class of the expected answer is later used to (a) filter out paragraphs that do not contain any word that can be cast in the same class as the expected answer, and (b) locate and extract the answers from the paragraphs. Finally, the answers are extracted and ranked based on their unification with the question.

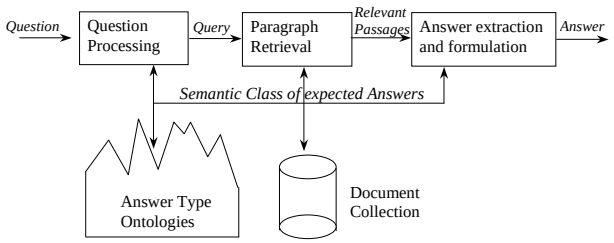

Figure 1. Architecture of a Q/A system.

2.1. Question Processing

To determine what a question asks about, several forms of information can be used. Since questions are expressed in natural language, sometimes their stems, e.g., *who*, *what* or *where* indicate the semantic class of the expected answer, i.e. PERSON, ORGANIZATION or LOCATION, respectively. To identify words that belong to such semantic classes, Name Entity Recognizers are used, since most of these words represent names. Name Entity (NE) recognition is a natural language technology that identifies names of people, organizations, locations and dates or monetary values.

However, most of the time the question stems are either ambiguous or they simply do not exist. For example, questions having *what* as their stem may ask about anything. In this case another word from the question needs to be used to determine the semantic class of the expected answer. In particular, the additional word is semantically classified against an ontology of semantic classes. To determine which word indicates the semantic class of the expected answer, the syntactic dependencies[1] between the question words may be employed [5, 10, 6].

1 Syntactic parsers publically available, e.g., [2, 4], can be used to capture the binary dependencies between the head of each phrase.

Sometimes the semantic class of the expected answers cannot be identified or is erroneously identified causing the selection of erroneous answers. The use of text classification aims to filter out the incorrect set of answers that Q/A systems provide.

2.2. Paragraph Retrieval

Once the question processing has chosen the relevant keywords of questions, some term expansion techniques are applied: all nouns and adjectives as well as morphological variations of nouns are inserted in a list. To find the morphological variations of the nouns, we used the CELEX [1] database. The list of expanded keywords is then used in the boolean version of the SMART system to retrieve paragraphs relevant to the target question. Paragraph retrieval is preferred over full document retrieval because (a) it is assumed that the answer is more likely to be found in a small text containing the question keywords and at least one other word that may be the exact answer; and (b) it is easier to process syntactically and semantically a small text window for unification with the question than processing a full document.

2.3. Answer Extraction

The procedure for answer extraction that we used is reported in [10], it has 3 steps:

Step 1) Identification of Relevant Sentences:
The Knowledge about the semantic class of the expected answer generates two cases: (a) When the semantic class of the expected answers is known, all sentences from each paragraph, that contain a word identified by the Named Entity recognizer as having the same semantic classes as the expected answers, are extracted. (b) The semantic class of the expected answer is not known, therefore all sentences, that contain at least one of the keywords used for paragraph retrieval, are selected.

Step 2) Sentence Ranking:
We compute the sentence ranks as a by product of sorting the selected sentences. To sort the sentences, we may use any sorting algorithm, e.g., the quicksort, given that we provide a comparison function between each pair of sentences. To learn the comparison function we use a simple neural network, namely, the perceptron, to compute a relative comparison between any two sentences. This score is computed by considering four different features for each sentence as explained in [10].

Step 3) Answer Extraction:
We select the top 5 ranked sentences and return them as answers. If we lead fewer than 5 sentences to select from, we return all of them.

Once the answers are extracted we can apply an additional filter based on text categories. The idea is to match the categories of the answers against those of the questions.

Next section addresses the problem of question and answer categorization.

3. Text and Question Categorization

To exploit category information for Q/A we categorize both answers and questions. For the former, we define as categories of an answer a the categories of the documents that contain a. For the latter, the problem is more critical since it is not clear what can be considered as categories of a question.

To define question categories we assume that users have in mind a specific domain when they formulate their requests (see [8]). The automatic models that we have study to classify questions and answers are: Rocchio [11] and SVM [13] classifiers. The former is a very efficient TC that can be used for real scenario applications. The latter is one of the best figure TC that provides *good* accuracy with a few training data.

3.1. Rocchio and SVM text classifiers

Rocchio and Support Vector Machines are both based on the Vector Space Model. In this approach, the document d is described as a vector $\vec{d} =< w^d_{f_1}, .., w^d_{f_{|F|}} >$ in a $|F|$-dimensional vector space, where F is the adopted set of features. The axes of the space, $f_1, .., f_{|F|} \in F$, are the features extracted from the training documents and the vector components $w^d_{f_j} \in \Re$ are weights that can be evaluated as described in [12].

The weighing methods that we adopted are based on the following quantities: M, the number of documents in the *training-set*, M_f, the number of documents in which the features f appears and l^d_f, the logarithm of the term frequency defined as:

$$l^d_f = \begin{cases} 0 & \text{if } o^d_f = 0 \\ log(o^d_f) + 1 & \text{otherwise} \end{cases} \quad (1)$$

where, o^d_f are the occurrences of the features f in the document d (TF of features f in document d).

Accordingly, the document weights is:

$$w^d_f = \frac{l^d_f \times IDF(f)}{\sqrt{\sum_{r \in F}(l^d_r \times IDF(r))^2}}$$

where the $IDF(f)$ (the Inverse Document Frequency) is defined as $log(\frac{M}{M_f})$.

Given a category C and a set of positive and negative examples, P and $\bar{P}$, Rocchio and SVM learning algorithms use the document vector representations to derive a hyperplane[2], $\vec{a} \times \vec{d} + b = 0$. This latter separates the documents that belong to C from those that do not belong to C in the *training-set*. More precisely, $\forall \vec{d}$ positive examples ($\vec{d} \in P$), $\vec{a} \times \vec{d} + b \geq 0$, otherwise ($\vec{d} \in \bar{P}$) $\vec{a} \times \vec{d} + b < 0$. $\vec{d}$ is the equation variable, while the gradient $\vec{a}$ and the constant b are determined by the target learning algorithm. Once the above parameters are available, it is possible to define the associated classification function, $\phi_c : D \rightarrow \{C, \emptyset\}$, from the set of documents D to the binary decision (i.e., belonging or not to C). Such decision function is described by the following equation:

$$\phi_c(d) = \begin{cases} C & \vec{a} \times \vec{d} + b \geq 0 \\ \emptyset & otherwise \end{cases} \quad (2)$$

Eq. 2 shows that a category is accepted only if the product $\vec{a} \times \vec{d}$ overcomes the threshold $-b$. Rocchio and SVM are characterized by the same decision function[3]. Their difference is the learning algorithm to evaluate the threshold b and the $\vec{a}$ parameters: the former uses a simple heuristic while the second solves an optimization problem.

3.1.1. Rocchio Learning The learning algorithm of the Rocchio text classifier is the simple application of the Rocchio's formula (Eq. 3) [11]. The parameters $\vec{a}$ is evaluated by the equation:

$$\vec{a}_f = \max\left\{0, \frac{1}{|P|}\sum_{d \in P} w^d_f - \frac{\rho}{|\bar{P}|}\sum_{d \in \bar{P}} w^d_f\right\} \quad (3)$$

where P is the set of training documents that belongs to C and ρ is a parameter that emphasizes the negative information. This latter can be estimated by picking-up the value that maximizes the classifier accuracy on a training subset called *evaluation-set*. A method, named the Parameterized Rocchio Classifier, to estimate *good* parameters has been given in [9].

The above learning algorithm is based on a simple and efficient heuristic but it does not ensure the best separation of the training documents. Consequently, the accuracy is lower than other TC algorithms.

3.1.2. Support Vector Machine learning The major advantage of SVM model is that the parameters $\vec{a}$ and b are evaluated applying the *Structural Risk Minimization principle* [13], stated in the statistical learning theory. This principle provides a bound for the error on the *test-set*. Such bound is minimized if the SVMs are chosen in a way that $|\vec{a}|$ is minimal. More precisely the parameters $\vec{a}$ and b are a solution of the following optimization problem:

$$\begin{cases} Minimize \quad |\vec{a}| \\ \vec{a} \times \vec{d} + b \geq 1 \quad \forall d \in P \\ \vec{a} \times \vec{d} + b < -1 \quad \forall d \in \bar{P} \end{cases} \quad (4)$$

It can be proven that the minimum $|\vec{a}|$ leads to a maximal margin[4] (i.e. distance) between negative and positive examples.

2 The product between vectors is the usual scalar product.

3 This is true only for linear SVM. In the polynomial version the decision function is a polynomial of support vectors.

4 The software to carry out both the learning and classifi cation algorithm for SVM is described in [7] and it can be downloaded from the web site *http://svmlight.joachims.org/*.

In summary, SVM provides a better accuracy than Rocchio but this latter is better suited for real applications.

3.2. Question categorization

In [9, 7], Rocchio and SVM text classifiers have been shown quite effective, thus, our idea is to adopt these models to classify questions. These latter should be considered as a particular case of documents, in which the number of words is rather small. This latter aspect implies that a big number of questions have to be used in the classifier training to reach a reliable statistical word distribution.

As in practical case such huge number of training questions is not available, we dropped the idea to learn the question categorization function only from questions. We have noticed that, when the training of document classifiers is applied, an explicit set of relevant words together with their weights is defined for each category (e.g. the vector $\vec{a}$). Our idea is to exploit Rocchio and SVM learning on category documents to derive question categorization functions.

We define for each question q a vector $\vec{q} = <w_1^q, .., w_{|F_q|}^q>$, where $w_i^q \in \Re$ are the weights associated to the question features in the feature set F_q, e.g. the set of question words. Then, we evaluate four different methods computing the weights of question features, which in turn determine five models of question categorization:

Method 1: We use l_f^q, the logarithm (evaluated similarly to Eq. 1) of the word frequency f in the questions q, together with the IDF derived from training documents as follows:

$$w_f^q = \frac{l_f^q \times IDF(f)}{\sqrt{\sum_{r \in F_q} (l_r^q \times IDF(r))^2}} \quad (5)$$

This weighting mechanism uses the Inverse Document Frequency (IDF) of features instead of computing the *Inverse Question Frequency*. The rationale is that question word statistics can be estimated from the word document distributions. When this method is applied to the Rocchio-based Text Categorization model, by substituting w_f^d with w_f^q we obtain a model call the $RTC0$ model. When it is applied to the SVM model, by substituting w_f^d with w_f^q, we call it SVM0.

Method 2: The weights of the question features are computed by the formula 5 employed in Method 1, but they are used in the Parameterized Rocchio Model [9]. This entails that ρ from formula 3 as well as the threshold b are chosen to maximize the categorization accuracy of the training questions. We call this model of categorization PRTC.

Method 3: The weights of the question features are computed by formula 5 employed in Method 1, but they are used in an extended SVM model, in which two additional conditions enhance the optimization problem expressed by Eq. 4. The two new conditions are:

$$\begin{cases} Minimize \quad |\vec{a}| \\ \vec{a} \times \vec{q} + b \geq 1 \quad \forall q \in P_q \\ \vec{a} \times \vec{q} + b < -1 \quad \forall q \in \bar{P}_q \end{cases} \quad (6)$$

where P_q and $\bar{P}_q$ are the set of positive and negative examples of training questions for the target category C. We call this question categorization model QSVM.

Method 4: We use the output of the basic Q/A system to assign a category to questions. Each question has associated up to five answer sentences. In turn, each of the answers is extracted from a document, which is categorized. The category of the question is chosen as the most frequent category of the answers. In case that more than one category has the maximal frequency, the set of categories with maximal frequency is returned. We named this ad-hoc question categorization method QATC (Q/A and TC based model).

4. Answers filtering and re-ranking based on Text Categorization

Many Q/A systems extract and rank answers successfully, without employing any TC information. For such systems, it is interesting to evaluate if TC information improves the ranking of the answers they generate. The question category can be used in two ways: (1) to re-rank the answers by pushing down in the list any answer that is labeled with a different category than the question; or (2) to simply eliminate answers labeled with categories different than the question category.

First, a basic Q/A system has to be trained on documents that are categorized (automatically or manually) in a predefined categorization scheme. Then, the target questions as well as the answers provided by the basic Q/A system are categorized. The answers receive the categorization directly from the categorization scheme, as they are extracted from categorized documents. The questions are categorized using one of the models described in the previous section. Two different impacts of question categorization on Q/A are possible:

- Answers that do not match at least one of the categories of the target questions are eliminated. In this case the precision of the system should increase if the question categorization models are enough accurate. The drawback is that some important answers could be lost because of categorization errors.
- Answers that do not match the target questions (as before) get lowered ranks. For example, if the first answer has categories different from the target question, it could shift to the last position in case of all other answers have (at least) one category in common with the question. In any case, all questions will be shown to the final users, preventing the lost of relevant answers.

An example of the answer elimination and answer reranking is given in the following. As basic Q/A system we

adopted the model described in Section 2. We trained[5] our basic Q/A system with the entire *Reuters-21578*[6]. In particular we adopted the collection Apté split. It includes 12,902 documents for 90 classes, with a fixed splitting between *test-set* and learning data (3,299 vs. 9,603). A description of some categories of this corpus is given in Table 1. Table

Category	Description
Acq	Acquisition of shares and companies
Earn	Earns derived by acquisitions or sells
Crude	Crude oil events: market, Opec decision,..
Grain	News about grain production
Trade	Trade between companies
Ship	Economic events that involve ships
Cocoa	Market and events related to Cocoa plants
Nat-gas	Natural Gas market
Veg-oil	Vegetal Oil market

Table 1. Description of some Reuters categories

2 shows the five answers generated (with their corresponding rank) by the basic Q/A system, for one example question. The categories of the document from which the answer was extracted is displayed in column 1. The question classification algorithm automatically assigned the $Crude$ category to the question.

The processing of the question identifies the word *say* as indicating the semantic class of the expected answer and for paragraph retrieval it used the keywords $k_1 = Director$, $k_2 = General$, $k_3 = energy$, $k_4 = floating$, $k_5 = production$ and $k_6 = plants$ as well as all morphological variations for the nouns. For each answer from Table 2, we have underlined the words matched against the keywords and emphasized the word matched in the class of the expected answer, whenever such a word was recognized (e.g., for answers 1 and 3 only). For example, the first answer was extracted because words producers, product and directorate general could be matched against the keywords production, Director and General from the question and moreover, the word *said* has the same semantic class as the word *say*, which indicates the semantic class of the expected answer.

The ambiguity of the word plants cause the basic Q/A system to rank the answers related to *Cocoa* and *Grain* plantations higher than the correct answer, which is ranked as the third one. If the answer re-ranking or elimination methods are adopted, the correct answer reaches the top as it was assigned the same category as the question, namely the *Crude* category.

5 We could not use the TREC conference data-set because texts and questions are not categorized.

6 Available at `http://kdd.ics.uci.edu/databases/reuters21578/`.

Next section describes in detail our experiments to prove that question categorization add some important information to select relevant answers.

5. Experiments

The aim of the experiments is to prove that category information used as described in the previous section is useful for Q/A systems. For this purpose we have to show that the performance of a basic Q/A system is improved when the question classification is adopted. To implement our Q/A and filtering system we used: (1) A state of the art Q/A system: improving low accurate systems is not enough to prove that TC is useful for Q/A. The basic Q/A system that we employed is based on the architecture described in [10], which is the current *state-of-the-art*. (2) The Reuters collection of categorized documents on which training our basic Q/A system. (3) A set of questions categorized according to the Reuters categories. A portion of this set is used to train PRTC and QSVM models, the other disjoint portion is used to measure the performance of the Q/A systems.

Next section, describes the technique used to produce the question corpus.

5.1. Question set generation

The idea of PRTC and QSVM models is to exploit a set of questions for each category to improve the learning of the PRC and SVM classifiers. Given the complexity of producing any single question, we decided to test our algorithms on only 5 categories. We chose Acq, $Earn$, $Crude$, $Grain$, $Trade$ and $Ship$ categories since for them is available the largest number of training documents. To generate questions we randomly selected a number of documents from each category, then we tried to formulate questions related to the pairs <document, category>. Three cases were found: (a) The document does not contain general questions about the target category. (b) The document suggests general questions, in this case some of the question words that are in the answers are replaced with synonyms to formulate a new (more general) question. (c) The document suggests general questions that are not related to the target category. We add these questions in our data-set associated with their true categories.

Table 3 lists a sample of the questions we derived from the target set of categories. It is worth noting that we included short queries also to maintain general our experimental set-up.

We generated 120 questions and we used 60 for the learning and the other 60 for testing. To measure the impact that TC has on Q/A, we first evaluated the question categorization models presented in Section 3.1. Then we compared the performance of the basic Q/A system with the extended Q/A systems that adopt the answer elimination and re-ranking methods.

Rank	Category	Question: *What did the Director General* say *about the energy floating production plants?*
1	*Cocoa*	” Leading cocoa producers are trying to protect their market from our product , ” *said* a spokesman for Indonesia ’s directorate general of plantations.
2	*Grain*	Hideo Maki , Director General of the ministry ’s Economic Affairs Bureau , *quoted* Lyng as telling Agriculture Minister Mutsuki Kato that the removal of import restrictions would help Japan as well as the United States.
3	*Crude*	Director General of Mineral and Energy Affairs Louw Alberts announced the strike earlier but *said* it was uneconomic .
4	*Veg-oil*	Norbert Tanghe, head of division of the Commission’s Directorate General for Agriculture, told the 8th Antwerp Oils and Fats Contact Days ” the Commission firmly believes that the sacrifices which would be undergone by Community producers in the oils and fats sector...
5	*Nat-gas*	Youcef Yousfi , director - general of Sonatrach , the Algerian state petroleum agency , indicated in a television interview in Algiers that such imports.

Table 2. Example of question labeled in the $Crude$ category and its five answers.

Acq	Which strategy aimed activities on core businesses? How could the transpacific telephone cable between the U.S. and Japan contribute to forming a join venture?
Earn	What was the most significant factor for the lack of the distribution of assets? What do analysts think about public companies?
Crude	What is Kuwait known for? What supply does Venezuela give to another oil producer?
Grain	Why do certain exporters fear that China may renounce its contract? Why did men in port’s grain sector stop work?
Trade	How did the trade surplus and the reserves weaken Taiwan’s position? What are Spain’s plans for reaching European Community export level?
Ship	When did the strikes start in the ship sector? Who attacked the Saudi Arabian supertanker in the United Arab Emirates sea?

Table 3. Some training/testing Questions

5.2. Performance Measurements

In sections 3 and 4 we have introduced several models. From the point of view of the accuracy, we can divided them in two categories: the (document and question) categorization models and the Q/A models. The former are usually measured by using Precision, Recall, and f-measure [14]; note that questions can be considered as small documents. The latter often provide as output a list of ranked answers. In this case, a *good* measure of the system performance should take into account the order of the correct and incorrect questions.

One method employed in TREC is the reciprocal value of the rank (RAR) of the highest-ranked correct answer generated by the Q/A system. Its value is 1 if the first answer is correct, 0.5 if the second answer is correct but not the first one, 0.33 when the correct answer was on the third position, 0.25 if the fourth answer was correct, and 0.1 when the fifth answer was correct and so on. If none of the answers are corrects, RAR=0. The Mean Reciprocal Answer Rank (MRAR) is used to compute the overall performance of Q/A systems[7], defined as $MRAR = \frac{1}{n}\sum_i \frac{1}{rank_i}$, where n is the number of questions and $rank_i$ is the rank of the answer i.

Since we believe that TC information is meaningful to prefer out incorrect answers, we defined a second measure to evaluate Q/A. For this purpose we designed the Signed Reciprocal Answer Rank (SRAR), which is defined as $\frac{1}{n}\sum_{j\in A}\frac{1}{srank_j}$, where A is the set of answers given for the *test-set* questions, $|srank_j|$ is the rank position of the answer j and $srank_j$ is positive if j is correct and negative if it is not correct. The SRAR can be evaluated over a set of questions as well as over only one question. SRAR for a single question is 0 only if none answer was provided for it.

For example, given the answer ranking of Table 2 and considering that we have just one question for testing, the MRAR score is 0.33 while the SRAR is -1 -.5 +.33 -.25 -.1 = -1.52. If the answer re-ranking is adopted the MRAR improve to 1 and the SRAR becomes +1 -.5 -.33 -.25 -.1 = -.18. The answer elimination produces a MRAR and a SRAR of 1.

5.3. Evaluation of Question Categorization

Table 4 lists the performance of question categorization for each of the models described in Section 3.1. We noticed better results when the PRTC and QSVM models were used. In the overall, we find that the performance of question categorization is not as good as the one obtained for TC in [9].

7 The same measure was used in all TREC Q/A evaluations.

	RTC0 f_1	SVM0 f_1	PRTC f_1	QSVM f_1	QATC f_1
acq	18.19	54.02	62.50	56.00	46.15
crude	33.33	54.05	53.33	66.67	66.67
earn	0.00	55.32	40.00	13.00	26.67
grain	50.00	52.17	75.00	66.67	50.00
ship	80.00	47.06	75.00	90.00	85.71
trade	40.00	57.13	66.67	58.34	45.45

Table 4. f_1 performances of question categorization.

5.4. Evaluation of Question Answering

To evaluate the impact of our filtering methods on Q/A we first scored the answers of a basic Q/A system for the test set, by using both the MRAR and the SRAR measures. Additionally, we evaluated (1) the MRAR when answers were re-ranked based on question and answer category information; and (2) the SRAR in the case when answers extracted from documents with different categories were eliminated. Rows 1 and 2 of Table 5 report the MRAR and SRAR performances of the basic Q/A. Column 2,3,4,5 and 6 show the MRAR and SRAR accuracies (rows 4 and 5) of Q/A systems that eliminate or re-rank the answer by using the RTC0, SVM0, PRTC, QSVM and QATC question categorization models.

The basic Q/A results show that answering the Reuters based questions is a quite difficult task[8] as the MRAR is 66.19%, about 15 percent points under the best system result obtained in the 2003 TREC competition. Note that the basic Q/A system, employed in these experiments, uses the same techniques adopted by the best figure Q/A system of TREC 2003.

The quality of the Q/A results is strongly affected by the question classification accuracy. In fact, RTC0 and QATC that have the lowest classification f_1 (see Table 4) produce very low MRAR (i.e. 62.24% and 60.70%) and SRAR (i.e. -18.94% and -31.99%). When the best question classification model QSVM is used, the basic Q/A performance improves with respect to both the MRAR (66.35% vs 66.19%) and the SRAR (-7.66% vs -37.24%) scores.

In order to study how the number of answers impacts the accuracy of the proposed models, we have evaluated the MRAR and the SRAR score varying the maximum number of answers, provided by the basic Q/A system. We adopted as filtering policy the answer re-ranking.

8 Past TREC competition results have shown that Q/A performances strongly depend on the questions/domains used for the evaluation. For example, the more advanced systems of 2001 performed lower than the systems of 1999 as they were evaluate on a more diffi cult test-set.

MRAR (basic)	.6619				
SRAR (basic)	-.3724				
Quest. Categ.	RTC0	SVM0	PRTC	QSVM	QATC
MRAR (re-rank.)	.6224	.6490	.6577	.6635	.6070
SRAR (elimin.)	-.1894	-.1349	-.0356	-.0766	-.3199

Table 5. Performance comparisons between basic Q/A and Q/A using answer re-ranking or elimination policies.

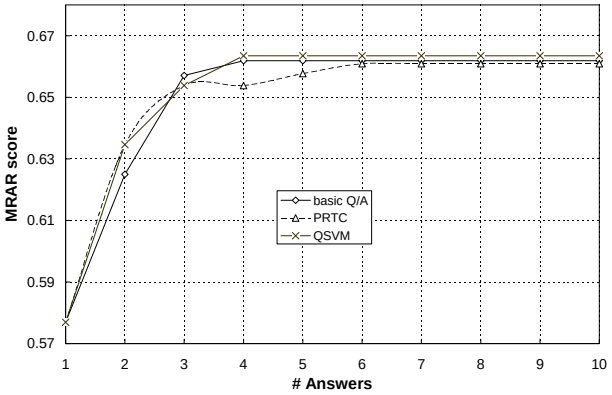

Figure 2. The $MRAR$ results for basic Q/A and Q/A with answer re-ranking based on question categorization via the PRTC and QSVM models.

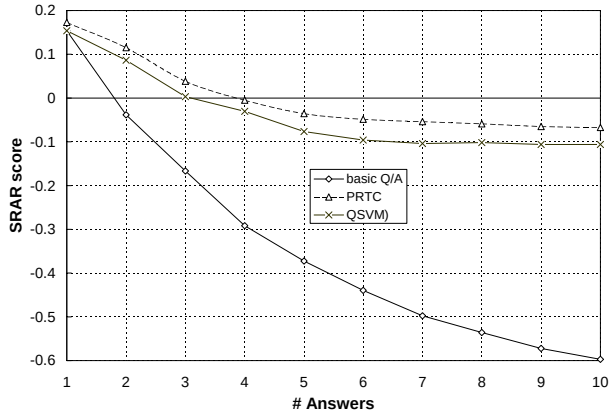

Figure 3. The $SRAR$ results for basic Q/A and Q/A with answer re-ranking based on question categorization via the PRTC and QSVM models.

Figure 2 shows that as the number of answers increases the MRAR score for QSVM, PRTC and the basic Q/A increases, for the first four answers and it reaches a plateau

afterwards. We also notice that the QSVM outperforms both PRTC and the basic Q/A. This figure also shows that question categorization per se does not greatly impact the MRAR score of Q/A.

Figure 3 illustrates the SRAR curves by considering the answer elimination policy. The figure clearly shows that the QSVM and PRTC models for question categorization determine a higher SRAR score, thus indicating that fewer irrelevant answers are left. Figure 3 shows that question categorization can greatly improve the quality of Q/A when irrelevant answers are considered. It also shows that perhaps, when evaluating Q/A systems with the MRAR scoring method, the "optimistic" view of Q/A is taken, in which erroneous results are ignored for the sake of emphasizing that an answer was obtained after all, even if it was ranked below several incorrect answers.

In contrast, the SRAR score that we have described in Section 5.2 produce a "harsher" score, in which errors are given the same weight as the correct results, but affect negatively the overall score. This explains why, even for a baseline Q/A, we obtained a negative score, as illustrated in Table 5. This shows that the Q/A system generates more erroneous answers then correct answers. If only the MRAR scores would be considered we may assess that TC does not bring significant information to Q/A for precision enhancement by re-ranking answers. However, the results obtained with the SRAR scoring scheme, indicate that text categorization impacts on Q/A results, by eliminating incorrect answers. We plan to further study the question categorization methods and empirically find which weighting scheme is ideal.

6. Conclusions

Question/Answering and Text Categorization have been, traditionally, applied separately, even if category information should be used to improve the answer searching. In this paper, it has been, firstly, presented a Question Answering system that exploits the category information. The methods that we have designed are based on the matching between the question and the answer categories. Depending on positive or negative matching two strategies allow to affect the Q/A performances: answer re-ranking and answer elimination.

We have studied five question categorization models based on two traditional TC approaches: Rocchio and Support Vector Machines. Their evaluation confirms the difficulty of automated question categorization as the accuracies are lower than those reachable for document categorization.

The impact of question classification in Q/A has been evaluated using the MRAR and the SRAR scores. When the SRAR, which considers the number of incorrect answers, is used to evaluate the enhanced Q/A system as well as the basic Q/A system, the results show a great improvement.

Acknowledgements

I would like to thank prof. Sanda Harabagiu for her support and comments. Without her expertise in Q/A systems this paper would not have been realized. Many thanks also are devoted to prof. Roberto Basili for his invaluable suggestions.

References

[1] R. H. Baayen, R. Piepenbrock, and L. Gulikers, editors. *The CELEX Lexical Database (Release 2) [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, 1995.

[2] E. Charniak. A maximum-entropy-inspired parser. In *In Proceedings of the 1st Meeting of the North American Chapter of the ACL*, pages 132–139, 2000.

[3] P. Clark, J. Thompson, and B. Porter. A knowledge-based approach to question-answering. In *proceeding of AAAI'99 Fall Symposium on Question-Answering Systems. AAAI*, 1999.

[4] M. Collins. Three generative, lexicalized models for statistical parsing. In *Proceedings of the ACL and EACLinguistics*, pages 16–23, Somerset, New Jersey, 1997.

[5] S. Harabagiu, M. Pasca, and S. Maiorano. Experiments with open-domain textual question answering. In *Proceedings of the COLING-2000*, 2000.

[6] S. M. Harabagiu, D. I. Moldovan, M. Pasca, R. Mihalcea, M. Surdeanu, R. C. Bunescu, R. Girju, V. Rus, and P. Morarescu. The role of lexico-semantic feedback in open-domain textual question-answering. In *Meeting of the ACL*, pages 274–281, 2001.

[7] T. Joachims. T. joachims, making large-scale svm learning practical. In B. Schlkopf, C. Burges, and M.-P. A. Smola (ed.), editors, *Advances in Kernel Methods - Support Vector Learning*, 1999.

[8] A. Moschitti. *Natural Language Processing and Automated Text Categorization: a study on the reciprocal beneficial interactions*. PhD thesis, Computer Science Department, Univ. of Rome "Tor Vergata", 2003.

[9] A. Moschitti. A study on optimal parameter tuning for Rocchio text classifier. In F. Sebastiani, editor, *Proceedings of ECIR-03, 25th European Conference on Information Retrieval*, Pisa, IT, 2003. Springer Verlag.

[10] M. A. Pasca and S. M. Harabagiu. High performance question/answering. In *Proceedings ACM SIGIR 2001*, pages 366–374. ACM Press, 2001.

[11] J. Rocchio. *Relevance feedback in information retrieval*. In G. Salton, editor, The SMART Retrieval System–Experiments in Automatic Document Processing, pages 313-323 Englewood Cliffs, NJ, Prentice Hall, Inc., 1971.

[12] G. Salton. *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley, 1989.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[14] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval Journal*, 1999.