Question Classification for Email*

Rachel Cotterill
University of Sheffield
r.cotterill@sheffield.ac.uk

Abstract

Question classifiers are used within Question Answering to predict the expected answer type for a given question. This paper describes the first steps towards applying a similar methodology to identifying question classes in dialogue contexts, beginning with a study of questions drawn from the Enron email corpus. Human-annotated data is used as a gold standard for assessing the output from an existing, open-source question classifier (QA-SYS). Problem areas are identified and potential solutions discussed.

1 Introduction and Motivation

In information retrieval, question classification is an important first stage of the question answering task. A question classification module typically takes a question as input and returns the class of answer which is anticipated. In an IR context, this enables candidate answers to be identified within a set of documents, and further methods can then be applied to find the most likely candidate.

The present work is motivated by a desire to identify questions and their answers in the context of written dialogue such as email, with the goal of improving inbox management and search. Reconstruction of meaning in a single email may often be impossible without reference to earlier messages in the thread, and automated systems are not yet equipped to deal with this distribution of meaning, as text mining techniques developed from document-based corpora such as newswire do not translate naturally into the dialogue-based world of email. Take the following hypothetical exchange:

From: john@example.com **To:** jane@example.com
`Jane,`
`Can you let me know the name of your lawyer?  Thanks.`
`John`

From: jane@example.com **To:** john@example.com
`Ally McBeal.`
`-- Jane`

This is an extreme example, but it serves to illustrate the "separate document problem" in email processing. Context is critical to pragmatic analysis, but with email and related media the context (and consequently, a single piece of information) may be spread across more than one document. In this case the second message in isolation gives no information concerning "Ally McBeal" as we do not have any context to put her in. However, by considering the question and answer pair together, we can discover that she is a lawyer (or, at the very least, that Jane believes or claims that to be the case; a philosophical distinction best left aside for the time being).

It is anticipated that questions in a dialogue context will exhibit a somewhat different range of types to typical IR questions, but that some will indeed be seeking the kind of factual information for which QA classifiers are currently designed. If this subset of fact-seeking questions can be reliably identified

*The author would like to thank GCHQ for supporting this research.

by an automated process, then existing question classifiers could be used to identify the expected answer type. Candidate answers could then be sought within the text of replies to the message in which the question is asked.

This paper briefly describes the gold standard data (Section 2), compares human annotation to the output of Ng & Kan's (2010) QA-SYS question classifier (Section 3), and proposes some future directions for research (Section 4).

2 The Data

In order to investigate question types in email, a suitable set of questions was required. To this end questions were automatically extracted from CMU's deduplicated copy of the Enron corpus (Klimt & Yang 2004). Of the 39,530 unique question strings identified in Enron outboxes, a random sample of 1147 were manually examined and annotated with the expected question type.

A number of taxonomies have been proposed for classifying answer types, of which Li & Roth's (2002) two-tier hierarchy is a reasonably comprehensive and widely-adopted example. Their coarse classes are Abbreviation (ABBR), Description (DESC), Entity (ENTY), Human (HUM), Location (LOC), and Numeric (NUM), and they then define a set of 50 subclasses. Table 1 shows how Li & Roth's taxonomy was mapped to the category labels adopted for the current work.

Cotterill 2010	**Li & Roth 2002**	**%**
Person(s)	HUM{individual,title}	2.53
Group or Organisation	HUM{group}	0.17
Descriptive text	HUM{description} DESC{manner, definition, description}	11.51
Reason	DESC{reason}	1.57
Date or Time	NUM{date, period}	3.57
Numeric	NUM{weight, volume/size, ordinal, percentage, count, speed, money, temperature, distance, other}	1.92
Phone	NUM{code}[1]	0.40
URL		0.17
Email		0.17
Place	LOC{country, state, city, mountain, other}	0.96
Animal	ENTY{animal}	0.00
Physical Object	ENTY{instrument, plant, body part, vehicle, food, product, substance}	0.30
Concept	ENTY{language, religion, letter, color, creative/artwork, disease/medical, currency}	0.40
Event or Activity	ENTY{event, sport, technique/method}	0.87
Other	ENTY{symbol, term, word, other} ABBR{abbreviation, expression}	0.00
Yes/No		41.33
Action Request		8.98
Rhetorical		5.23
Multiple		3.23
Non-Question		16.74

Table 1: The new dialogue taxonomy, with mappings to Li & Roth where applicable, and percentage distribution in the Enron sample

[1]Phone number is actually a subset of the NUM:code category, but it accounts for all instances in the Enron sample.

"Are you guys still thinking of maybe joining us skiing?"	Yes/No
"Did you know Moller was going to be on TV or were you just channel surfing?" "Do you stock those wheels and tires or would I have to order them?"	Multiple choice
"Will it ever end???"	Rhetorical
"Would you please handle this?" "Also, could you check for reservations at the Georgian hotel in Santa Monica?"	Action Request

Table 2: Examples of questions in dialogue-specific categories

A number of extra categories were added to account for the nature of the data, as identified by preliminary experiments. Examples of questions falling into some of the new categories are presented in Table 2.

It is important to observe that a massive 75.5% of questions in the Enron sample do not fall into any of the categories defined by Li & Roth. Assuming that this is a fair representation of the distribution across the Enron corpus (if not email as a whole) then we are clearly justified in stating that some further work will be required before question classification can be meaningfully applied to the email task.

The most common category is Yes/No, giving a "most common class" baseline of 41.3%. That is to say, a classification system which classified every question as a Yes/No question would expect to see accuracy in this region, and any results must be considered in this context.

The most common of the IR-derived categories is Description, representing 11.51% of questions overall, or 46.2% of those falling into IR categories. This compares to 26.6% reported across the equivalent categories in Li & Roth's analysis of TREC questions.

Full details of the Enron question dataset will be published in due course.

3 Performance of QA-SYS

QANUS (Ng & Kan 2010) is an open-source question answering framework which uses the Li & Roth categories in its question classification module. The framework is designed to be extensible, which makes it a good candidate for further work. However, the results presented in this section deal only with the output of the QA-SYS default question processing module as supplied with QANUS v26Jan2010. The question classification component of QA-SYS is an instance of the Stanford classifier, a supervised learning module trained on a dataset of information retrieval questions.

Ng & Kan do not report their question classification accuracy, providing figures only for the "factoid accuracy" of the end-to-end question answering system, which makes it difficult to compare their results to the present study. However Huang, Thint & Cellikyilmaz (2009) publish results for a maximum entropy classifier trained on similar IR data, reporting an encouragingly high accuracy of 89.0%.

QA-SYS question classification was used to provide an automatic classification for each of the questions extracted from the Enron dataset. In order to assess the performance of the system, the results were compared to the hand-annotated examples.

QA-SYS output agreed with human annotation in only 13.4% of cases overall – much lower than the "most common class" baseline defined above. However, this figure is artificially low as QA-SYS supplies a result in all circumstances, without any associated level of confidence. The system will therefore provide an incorrect result in cases where it does not have an appropriate category (even when faced with a nonsense string).

This may be acceptable behaviour within information retrieval, particularly for participating in competitions when there is a high expectation of the question falling into one of Li & Roth's categories, but for dialogue questions it produces a number of undesirable effects. Any competent end-to-end system would need (at a minimum) to filter out nonsense strings, and direct questions to appropriate classifiers

“Remind me when your wedding date is?”	NUM:date	DateTime
“Also, who is following up on the VA license?”	HUM:ind	Person
“What is our strategy/plan in agricultural commodities training?”	DESC:desc	Description

Table 3: Examples of questions correctly classified

Category	Recall	Precision	F-measure
Description	62.8	28.6	39.3
DateTime	53.7	28.6	37.3
Numeric	40.9	15.5	22.5
Reason	61.1	10.2	17.5
Person	65.5	6.7	12.2
Place	45.5	5.2	9.3
Event	10.0	2.9	4.6

Table 4: Recall and precision by category

based on language (therefore removing the need to attempt an intelligent classification of texts in multiple languages). Considering the proportion of questions in our sample which fell into the new categories of our extended taxonomy, the framework should also be extended to include a number of classifiers to handle these data types specifically.

We are therefore justified in considering what might happen if a pre-classifier fed to QA-SYS only those questions which it may stand some chance of categorising correctly. Including only those questions falling into categories on which QA-SYS has been trained, output agrees with human annotation in 55.0% of cases. Table 3 presents a small number of examples where the QA-SYS annotation agreed with human assessment.

It may also be instructive to consider the recall and precision on a per-category basis, as there is a strong variation between the success rates for different QA-SYS categories. Table 4 gives the figures for those classes with at least 10 examples in the current dataset, and which QA-SYS claims to address.

This shows that some categories with the highest recall (e.g. Person, Reason) suffer from low precision, but examination of the full confusion matrix shows that the incorrect categorisation is largely accounted for by the categories for which QA-SYS is not trained (particularly Yes/No questions). If reliable classifiers could be trained to filter out these question types at an earlier stage, the validity of QA-SYS results would be significantly improved.

However, there are some features of QA-SYS question classification which cannot be resolved by simply adding additional categories to the classifier framework.

Most notably, the system exhibits a high degree of case sensitivity. For example, the two strings “What do you think?” and “what do you think?” are both present in the Enron corpus. To a human eye the lack of capitalisation is unlikely to affect the meaning, but QA-SYS categorises these two sentences differently: the former as DESC:desc, the latter as ENTY:term.

A further example of case-sensitivity is found in the response of QA-SYS to questions written entirely in uppercase. Of the eleven examples in the dataset which contain only uppercase letters, all are classified as ABBR:exp. The ‘uppercase’ feature seems to overwhelm any other considerations (such as question word) which may be present. For instance “WHAT?” is classified as ABBR:exp, whereas “What?” and “what?” are (correctly) classified as DESC:desc.

Certain words also have a significant impact on the classification, regardless of the syntax of the question. For example, a question containing the word ‘percent’ is likely to be classified as NUM:perc, a question containing the word ‘week’ is likely to be classified as NUM:date, and a question containing the word ‘state’ is likely to be classifed as some subtype of LOCATION.

Other lexical effects were surprising by their absence. For instance, of 111 questions (in the entire Enron question-set) beginning “What time…” only eleven are classified as requiring the NUM:date response.

“How many kids are in the class and who is the instructor?”	NUM:count
“do you want to get together on friday or saturday and where?”	LOC:other
“How (and when) do you plan to get there?”	DESC:manner

Table 5: Examples of compound questions

Another small but important set of questions, which are barely represented in the current dataset, are compound questions. These are cases, such as the examples in Table 5, in which more than one answer is expected. In all of these examples, the category generated by QA-SYS can hardly be called incorrect, however it is not the whole story. Presently QA-SYS does not allow for multiple answer types. This is worthy of further study.

4 Future Work

The present work should be extended using a larger dataset to train additional classifiers for the answer types which are beyond the scope of IR classifiers such as QA-SYS. A larger dataset will also enable further analysis, for example to identify any common features of questions which prove particularly hard to categorise. Specific work to identify further examples in the very small categories (including a representative sample of compound questions) would also be beneficial.

The next step is to extend the QANUS framework with additional classifiers trained on Enron data, and this work should be thoroughly tested to ensure it is not over-fitted to Enron. There is a wealth of public dialogue data on the web, available from textual media such as web forums and Twitter, which may be reasonably expected to have some characteristics in common with email and which could be used for testing the classifiers.

Recent work on email has considered the task of highlighting messages within an inbox which require action (e.g Bennett & Carbonell 2005, achieving 81.7% accuracy). This is an interesting result for us as the set of actions intersects with the set of questions: some questions have the pragmatic force of an action request. It would be interesting to examine the size of this intersection.

5 References

Bennett, Paul and Carbonell, Jaime. 2005. ‘Detecting Action-Items in Email’ in proceedings of *SIGIR ‘05*.

Huang, Zhiheng, Thint, Marcus, and Celikyilmaz, Asli. 2009. ‘Investigation of Question Classifier in Question Answering’ in proceedings of *EMNLP ‘09*.

Klimt, Bryan and Yang, Yiming. 2004. ‘Introducing the Enron Corpus’ in proceedings of *First Conference on Email and Anti-Spam*.

Li, Xin and Roth, Dan. 2002. ‘Learning Question Classifiers’ in proceedings of *COLING 2002*.

Ng, Jun-Ping and Kan, Min-Yen. 2010. *QANUS: An Open-source Question-Answering Platform*, from http://www.comp.nus.edu.sg/∼junping/docs/qanus.pdf