A Study on Expert Sourcing Enterprise Question Collection and Classification

Yuan Luo[1], Thomas Boucher[2], Tolga Oral[3], David Osofsky[3], Sara Weber[3]
[1]MIT CSAIL, 32 Vassar St, Cambridge, MA, 02139
[2]UMass CS, 140 Governors Drive, University of Massachusetts, Amherst, MA, 01003
[3]IBM CIOLab, 1 Rogers St, Cambridge, MA, 02139
E-mail: yuanluo@mit.edu, boucher@cs.umass.edu, {tolga_oral,david_osofsky,sara_weber}@us.ibm.com

Abstract

Large enterprises, such as IBM, accumulate petabytes of free-text data within their organizations. To mine this big data, a critical ability is to enable meaningful question answering beyond keywords search. In this paper, we present a study on the characteristics and classification of IBM sales questions. The characteristics are analyzed both semantically and syntactically, from where a question classification guideline evolves. We adopted an enterprise level expert sourcing approach to gather questions, annotate questions based on the guideline and manage the quality of annotations via enhanced inter-annotator agreement analysis. We developed a question feature extraction system and experimented with rule-based, statistical and hybrid question classifiers. We share our annotated corpus of questions and report our experimental results. Statistical classifiers separately based on n-grams and hand-crafted rule features give reasonable macro-f1 scores at 61.7% and 63.1% respectively. Rule based classifier gives a macro-f1 at 77.1%. The hybrid classifier with n-gram and rule features using a second guess model further improves the macro-f1 to 83.9%.

Keywords: question classification, expert sourcing, machine learning

1. Introduction

Watson Sales Assistant (WSA) is an internal pilot of adapting IBM Watson to help IBM salespeople get answers to their questions. In order to leverage its rich analytic components and parallel processing ability to digest petabytes of web repository content, we adapted the Watson platform to IBM enterprise content. Among others, adaptation of the question classification model is the first crucial step. Watson was trained on questions that look for named entities given declarative sentences as clues. On the other hand, WSA needs to answer sales questions of great variety, for which we had few existing archives, much less those with answers. We took an expert sourcing approach by inviting IBM Subject Matter Experts (SMEs) within the IBM sales division, whose jobs involve answering IBM salespeople's questions, to contribute to the set of questions and answers. The SMEs were asked to provide questions for which they have answers, as the first step in a boot strapping process that would allow us to train a question answering model with the collected corpus. To our best knowledge, there are few if any question classification corpora available for enterprise content question answering. To be broadly useful as a shared dataset, our question classification scheme is designed to be general and does not contain organization specific classes.

The sales questions can be interrogative questions with wh-words, or yes/no questions, or even imperative sentences. They may ask for a named entity, a description of a product, or reasoning on a fact. Such variety and complexity brings several challenges to the collection and classification task of sales questions. This paper describes the approach used to collect, annotate and automatically classify those sales questions.

2. Related Work

General domain question classification is an area of active research, since the ability to accurately answer a question depends on correct question classification (Hovy et al., 2001). In early TREC question answering tracks, rule-based approaches were typically applied with hand crafted heuristic rules (Voorhees, 1999; Voorhees and Tice, 2000; Voorhees, 2002). In (Hermjakob, 2001), the author augmented the rules with both semantic enrichment and an additional Penn Treebank questions. In (X. Li and Roth, 2002), the authors released the UIUC question classification dataset annotated using a two level taxonomy and classified questions with classes from each level, using the Sparse Network of Winnows (SNoW) algorithm. Many works have since focused on statistical models including SVM (Hacioglu and Ward, 2003; Solorio et al., 2004), log-linear models (Blunsom et al., 2006), Maximum Entropy models (M. L. Nguyen et al., 2007), kernel methods (Moschitti et al., 2007; Moschitti et al., 2011), among others (Pinto et al., 2002; Zhang and Lee, 2003). Only a few have continued rule-based approach, e.g., (Ray et al., 2010). Besides widely used n-grams, advanced semantic-syntactic features such as parse subtrees have been tested in statistical approaches with both positive results (M. L. Nguyen et al., 2007) and negative results (Moschitti et al., 2011). Most works used the UIUC corpus (X. Li and Roth, 2002) with minor variations. We also observed that a straightforward hybrid rule-based and statistical model could be beneficial, but has not yet been extensively investigated in previous studies.

3. Methods

During the initial stage of collecting the sales questions, we started drafting a question classification guideline by analyzing the first 600 questions collected. We then started the expert sourcing question collection and classification approach. With the feedback from annotators, our understanding of the characteristics of sales questions evolved, so did the classification guideline.

3.1 Expert Sourcing Approach for Question Collection

We invited a group of 630 SMEs to participate in the voluntary effort. Gifts were awarded to drive participation, at the levels of 25 submitted questions and answers (awarded a cap), 50 (a tumbler), and 100 (a backpack). A total of 165 SMEs contributed questions and answers.

Figure 1 shows that 72 SMEs (44%) contributed 90% of the initial set of questions and answers (the most productive participant contributing 5%). We tracked and published participation record to improve the quality of the questions and answers, as users' reputations were at stake.

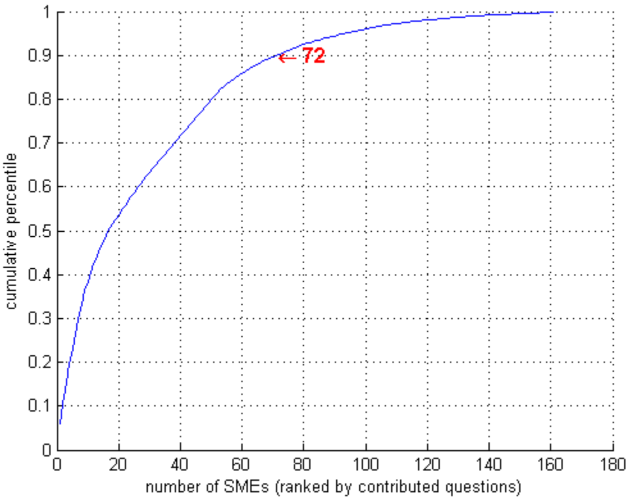


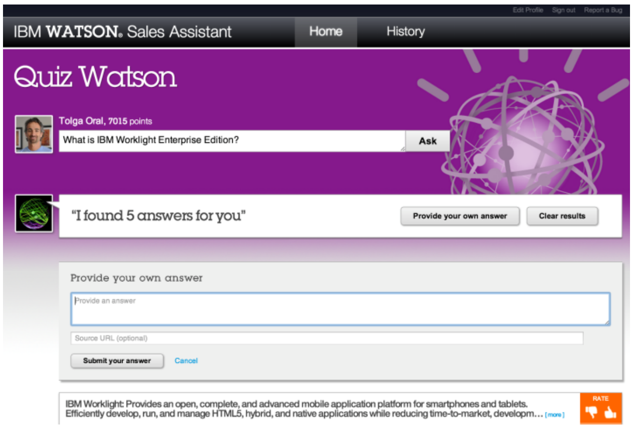
Figure 1: Cumulative Contribution by SMEs. The SMEs are ranked in descending order on number of questions contributed.

Figure 2 shows the question collection UI, integrated with the answering system. SMEs are asked to enter questions that might be asked by other salespeople and give the answers in the "Provide your own answer" textbox. In the textbox below, SMEs could optionally provide evidence URLs that are sources of their answers. During question analysis, the optionally provided evidence URLs were used to verify the correctness of the answers and also allowed discovery of new content sources to be included in the system. The phase 1 expert sourcing effort ran over a three month period and collected 3602 questions in total.

3.2 Question Classification along Semantic and Syntactic Axes

Following (Hovy et al., 2002; Garcia-Fernandez et al., 2010), we also take into account questions' syntactic characteristics in addition to what the questions are asking for (semantic). As enterprise repository is a different content from general web content, we built our classification scheme in a bottom-up and data-driven fashion, keeping in mind the implication on answer generation and ranking. Our semantic component shares some classes with the UIUC dataset, but also has noticeable differences. For example, we have no classes of "food", "animal" etc. On the other hand, our "approach" and "info" classes are heavily populated.

Currently 3602 questions have been manually classified with the following semantic categories: {named entity, degree, location, info, time, fact, definition, approach, reason, relation, difference}. Most categories are self-explanatory. The "info" questions ask for a URL or a reference. The "degree" questions ask for a numeric answer, or abstract degrees. The "named entity" questions ask for named entities other than "degree", "location", "info" and "time". If a question cannot be otherwise labeled, it is labeled as asking about a "fact".



Tolga Oral, 7015 points

What is IBM Worklight Enterprise Edition?

"I found 5 answers for you"

Provide an answer

Source URL (optional)

IBM Worklight: Provides an open, complete, and advanced mobile application platform for smartphones and tablets. Efficiently develop, run, and manage HTML5, hybrid, and native applications while reducing time-to-market, developm… [more]

Figure 2: Question and answer collection tool UI, integrated with the WSA answering system.

Questions of the same semantic categories may be found in different syntactic constructs, which we identified as: {who, what, which, when, where, how, polar, declarative, imperative}.

We define a question class as the combination of its semantic category and its syntactic construct[1]. Examples:

a. [*named entity/what*] What is IBM's Smarter Cities offering?
b. [*definition/what*] What is an application server?
c. [*fact/what*] What are the benefits of IBM Cognos 64-bit?
d. [*time/when*] When will IBM deliver CloudBurst 2.1?
e. [*time/what*] What is the target delivery date for IBM WebSphere 9.0?
f. [*approach/how*] How do I install IBM Smarter Analytics on Red Hat?

[1] Some combinations may not exist in reality, e.g., time/why.

g. [*approach/what*] What is the easiest way to install PureApplication System on Red Hat?
h. [*approach/polar*] Can I install InfoSphere Replication Server on Red Hat?
i. [*approach/declarative*] My client wants instructions for installing ISDM on Red Hat.
j. [*approach/imperative*] Tell me how to install Lotus Notes on Red Hat.

Semantic classes directly affect downstream steps in the question answering workflow. On the other hand, the syntactic class can also be useful for answer generation and scoring. For example, “time/what” questions may prefer a time format (e.g., what year, what month, or what date) while “time/when” questions generally do not have such restrictions. Another more involved example concerns an existing Watson answer evidence scorer that replaces the focus[2] of the question with candidate answers and then computes a matching score between the modified question text and relevant sentence returned from search. This scorer is directly applicable to “degree/what” while not directly applicable to “degree/how” as the latter often do not have a nominal focus. The recognition of the syntactic construct is nontrivial for computers, especially for sentences with complex clauses. Thus human annotations for training the classifier are necessary.

We also found that it is easier for human annotators to first look for a question’s syntactic class then determine semantic classes, since making annotators first classify syntactic construct (easier step) prunes the options for semantic goals (harder step). We point out that our classification has the flexibility to be folded to the semantic-only classification by dropping the syntactic component, if that is determined to be more appropriate.

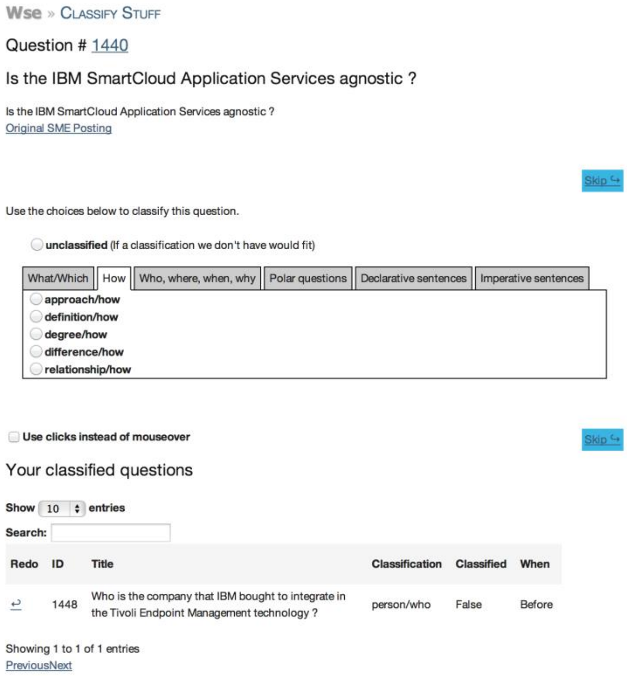
3.3 Expert Sourcing Classification and Enhanced Inter-Annotator Agreement

The collected questions were classified by twelve annotators who were members of the WSA adaptation team. We developed a classification annotation tool, which presents each question along with candidate question classes organized by their syntactic constructs. In Figure 3, each tab includes a list of possible question classes associated with certain syntactic construct. We supplied a classification guideline[3] with a description and examples of each question class.

A key feature of the annotation tool is the randomization of questions. By default, questions exported from the question answering tool tend to be grouped by the SME who often generates a series of questions. Randomization prevents a single annotator who might have difficulty understanding a particular SME’s questions from operating on a complete series of those questions.

We had group training and practice sessions for question classification, but classification of the 3602 questions using the annotation tool was performed individually. The tool enforces independent classification by preventing annotators from viewing classifications from others. If the first two classifications of a question match, the tool deems classification of that question complete and removes it from the queue of unclassified questions. Because questions are assigned to two of twelve annotators randomly, there are $\binom{12}{2} = 66$ combinations of annotators that are involved in initial inter-annotator agreement run.

If the first two annotators’ classifications of a question do not agree, the question remains in the unclassified queue. The tool only removes a question from the queue when at least 70% of the classifications for that question agree. This enhanced inter-annotator agreement guarantees a high level of annotation confidence. For example, if two annotators disagree on a classification, a third annotator alone cannot break the tie.



Question # 1440

Is the IBM SmartCloud Application Services agnostic ?

Is the IBM SmartCloud Application Services agnostic ?
Original SME Posting

Use the choices below to classify this question.

○ **unclassified** (If a classification we don't have would fit)

○ **approach/how**
○ **definition/how**
○ **degree/how**
○ **difference/how**
○ **relationship/how**

□ **Use clicks instead of mouseover**

Your classified questions

Show 10 **entries**
Search:

Redo	ID	Title	Classification	Classified	When
↩	1448	Who is the company that IBM bought to integrate in the Tivoli Endpoint Management technology ?	person/who	False	Before

Showing 1 to 1 of 1 entries
PreviousNext

Figure 3: Classification Tool UI. Syntactic classes are grouped into tabs to balance the list length of the associated semantic classes.

For the 394 questions with below the 70% enhanced agreement after exhausting 12 annotators, we have all involved annotators discuss together and see if they can resolve the disagreement. This eliminates common problems of mislabeling and misunderstanding of labels. However, there are still 126 (or 3.5% out of 3602) questions without agreement. We excluded these questions from our ground truth, trading slightly smaller coverage

[2] In Watson terminology, focus is the part of a question that is a reference to a named entity answer. For example, “what” is the focus in “What is the maximum memory capacity for a single node x3850 X5 server?” Note that focus is by definition nominal.

[3] An updated question classification guideline is submitted as supplemental material.

for better quality. For excluded questions, we categorized them by their majority annotations, as in Figure 4. The three most ambiguous categories are “fact/polar”, “fact/what”, and “named entity/what”. For example, some annotators are uncertain about either “fact/polar” or “approach/polar” for the question “Can I use CloudBurst to reduce delivery time of new services and offerings?” We refer the reader to the classification guideline for more discussion on ambiguity.

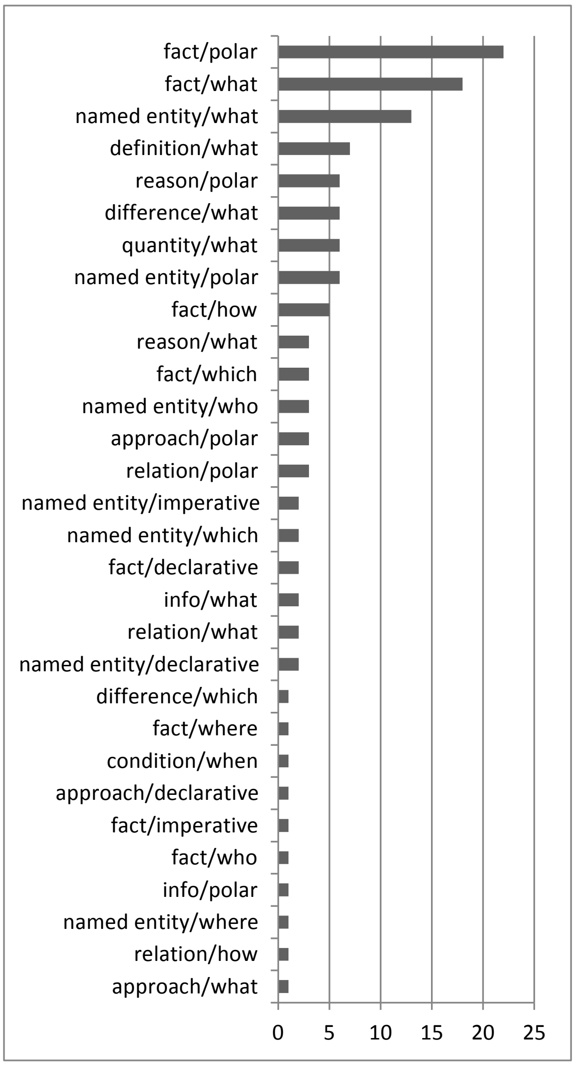


Figure 4: Percentages of questions not passing 70% agreement according to majority annotations

3.4 Data Preparation

Our dataset consists of those questions with at least 70% of classification agreement or with agreement after discussion. We only retained the question classes with 10 or more examples in the corpus. There are 3470 such questions in total. To perform classification task, we split our question set into a 70% training set and a 30% testing set, stratified by the question class[4]. The training set has 2420 questions in total. The testing set has 1050 questions in total. The distributions of question classes within training and testing sets are shown in Table 1. Note that the distribution of 22 classes is highly imbalanced with the three most populated classes being “named entity/what”, “definition/what”, and “fact/what”.

Classes	Train	Test	Classes	Train	Test
def/what	582	247	deg/what	43	20
ne/what	464	197	diff/what	30	14
fact/what	336	152	ne/polar	27	12
time/when	140	61	reason/why	26	10
fact/polar	139	60	fact/which	26	8
info/where	124	60	appr/what	19	8
ne/who	117	50	info/what	18	6
ne/which	103	46	info/polar	13	5
deg/how	65	30	appr/polar	12	7
appr/how	63	26	diff/how	10	5
loc/where	55	22	rel/how	8	4

Table 1: Question classes distribution, sorted by frequency. Classes are in the form of semantic/syntactic classification, see section 3.1. Abbreviations used in the table: approach (appr), relation (rel), location (loc), difference (diff), named entity (ne), definition (def), degree (deg).

3.5 Rule-Based Classifier

By analyzing the training question set, we developed a rule-based classifier that incorporates both syntactic and semantic features of sentences. We hypothesized that with carefully selected syntactic and semantic features, classification could be characterized as explicit and easy-to-understand rules. The classifier first launches individual recognizers each corresponding to one question class (e.g., a “definition/what” recognizer). Each recognizer extracts binary features (e.g., “subject is what”, “object is what”) by analyzing the question sentence. The classifier then pools the recognizer responses and makes collective decision on the final classification of the question. We next explain the steps in more detail.

3.5.1 Syntactic Feature Extraction

We ran the ESG Parser (McCord, 2010) to parse each sentence. The ESG Parser is a mixed syntactic and semantic parser in that the parser also performs Word Sense Disambiguation (WSD) in addition to identifying syntactic constructs. The parser is built on the concept of “slot” that captures the syntactic and sometimes semantic roles for phrases in a sentence. For our task, main subjects, objects and predicates are of particular interest. For example, in the info/polar question “Do you have references on what IBM Smarter Cities is?”, if the word “what” is taken as the main predicate and “IBM Smarter Cities” is taken as the main subject, the question will be wrongly classified as “definition/what”. To ensure the capture of main subjects and predicates, we test if their parent node is the top node in the ESG Parser parse tree. Ensuring the capture of main objects is more complicated. First, we in-

[4] Due to the randomness of the coin toss, the exact percentage may differ in each class, but still close to 7:3 ratio.

clude both verb objects and prepositional objects. For example, ESG Parser treats “what” as a prepositional object in “What does ISDM stand for?” Second, a main object may be demoted one level in the parse tree due to the presence of an auxiliary verb. In the above sentence, the node “what” is a child of the verb node “does”, which is then a child of the top node of the parse tree. There are other syntactic constructs that generate useful features to the recognizers, including left modifiers, right modifiers, nouns coordinated by conjunctions or symbols such as forward slash. In particular, when trying to identify whether a coordinated noun phrase such as “IBM Lotus Live and Infosphere” refer to product names, we traced the lconj and rconj slots of the coordinator “and”, checked whether either slot refers to product names, then aggregated the findings.

3.5.2 Semantic and Other Feature Extraction

To derive the semantic features, we partly rely on existing ontologies, taxonomies and lexical resources such as WordNet (Fellbaum, 1998), YAGO (Suchanek et al., 2007) and an in-house taxonomy (Murdock and Welty, 2006). We also add to the in-house taxonomy additional categories and terms that are mined from IBM repositories (e.g., Product and Computing Topic categories).

The recognizers iterate through an ESG parse in breadth first order, collecting the aforementioned syntactic and semantic features as they proceed. Note that extracting these features often requires collaboration between the ESG Parser and multiple ontologies. For example, the feature “subject is YAGO named instance” requires first identifying the main subject then testing on whether the content of the main subject is a named instance in the YAGO/WordNet ontology. In addition to existing ontologies, we compiled lists of words that may be cues for question classes. For example, *approach nouns* include “approach”, “way”, “process” etc, *product nouns* include “product”, “software”, “solution”, “asset” etc. Moreover, we use morphological features such as “subject has capitalization”, “predicate has capitalization[5]”, as well as morphologic-syntactic pattern such as “’what is’ followed by words that are each capitalized”, which is a strong indication that if matched, the sentence is a definition. We refer the reader to the supplemental material for a complete list of raw features.

3.5.3 Composing the Rules

A total of 120 handpicked features are incorporated into one or more rules. For example, in the definition/what recognizer, one simple rule looks like “subject is what && predicate has no article”, where we use the broad sense of articles that include “a/an”, “the”, “some”, and genitive cases such as “IBM’s.” Another rule involving both semantic and syntactic features looks like “predicate is what && (subject has capitalization || subject is YAGO named instance)”. One single rule is not intended to be comprehensive, but precision is emphasized. Each recognizer then tests its associated rules, and returns yes if one or more rules are satisfied. A typical recognizer usually has 5 or 6 firing rules. The classifier then treats all recognizers as a partially ordered set. For example, the classes “definition/what”, “named entity/what”, “fact/what” are preferred in that order. In contrast, the classes “definition/what” and “definition/which” can exchange their ranks. The intention of this design is to break the tie when multiple recognizers fire simultaneously. The most general semantic class with certain syntactic class (e.g., fact/what for all */what questions) captures questions for which no other recognizers are fired. We included all rules in the supplemental material.

3.6 Statistical Classifier

We trained and tested six different question classifier models, where each model relies on the same learning algorithm but uses features including n-grams, recognizer-associated rules or recognizer output (details in Section 4.2). All models used an L1-regularized multinomial logistic regression classifier to predict the question classes. We tried multiple standard machine learning algorithms, including Bayesian methods and support vector machines, and selected logistic regression because of its accuracy and speed on our data. The resulting probabilities were used as confidence measures of the question classifier which were used later in the Watson pipeline for candidate answer scoring. We experimented with three techniques for multinomial logistic regression, true multinomial, a one-vs-one strategy (one classifier per class pair), and a one-vs-all strategy (one classifier per class). Our experiment showed that a one-vs-all strategy performed best, while a true multinomial performed almost identically well and a one-vs-one strategy performed the weakest. To encourage sparsity and to prevent overfitting, we used an L1-regularized model. Using L1 instead of L2-regularized logistic regression has been shown to require fewer training examples to learn well (Ng, 2004), and experiment showed that the L1-regularized models performed better. All models reported in this paper were trained using a one-vs-all strategy with an L1 regularizer, and all model parameters were chosen using 10-fold cross-validation exclusively on the training data.

4. Experiments and Results

Our experiment has two parts. The first part analyzes inter-annotator agreement and the second part runs multiple classifiers on different feature configurations.

4.1 Inter-Annotator Agreement

We use the Kappa score (Fleiss, 1981) to measure the inter-annotator agreement. Kappa is thought to take into account agreement that could occur by chance hence is more robust than simple agreement percentages. In addition to Kappa score of the originally annotated questions, we also computed the Kappa score after disagreement resolving attempts. If there is an annotation class receiving more than 70% votes or the question disagreement is

[5] Besides the interrogative words “What”, “How” etc., of course.

resolved by discussion, we treat the question as having two agreed annotations. Otherwise, the question has disagreed annotations. The Kappa score before disagreement resolving is 0.700, which is considered as good agreement (Fleiss, 1981). Moreover, the Kappa score after disagreement resolving is 0.923, which demonstrates high inter-annotator agreement.

4.2 Classification Results

As a baseline comparison, we used a logistic regression model with a unigram and bigram feature set. Larger n-gram features did not improve the performance of the model and were omitted. When vectorizing the question corpus, we used neither stopwords nor minimum term frequency (TF). Since the questions asked to WSA were mostly terse, the model performed best with these unrestrictive settings. TF and TF-IDF normalization decreased the performance of our model and were also omitted.

We next experimented with the rule-based classifier and two logistic regression models using rule features, one using the 120 raw rule responses as features and the other using the 22 class recognizer responses as features. The difference between the two settings is that the recognizer response features have more human insights. The class recognizers and the raw rules are all binary features, since for each question, a class recognizer or a raw rule either fires or does not fire. We also compared two hybrid models, one using the raw rules and the n-gram features, and the other using the recognizer responses and the n-gram features.

Lastly, we chose the best performing model, the class recognizer and n-gram hybrid, and included the second guess results. Watson generated an analysis pipeline for each candidate answer, and the benefit of including the correct question class vastly outweighed the cost of generating a second analysis pipeline. To prevent the system from always providing a second guess, we used a thresholding approach. If the question class with the highest probability had a probability below a given threshold (0.6 in this work), then the classifier was allowed to give a second guess.

To comprehensively judge each configuration's performance, we computed the macro- and micro-averaged precision, recall and f1-score, as shown in Table 2. We performed significance tests using approximate randomization test (Noreen, 1989) comparing rule-based recognizers vs. n-grams, "1+3" hybrid vs. rule-based, and "1+3" hybrid with second guess vs. "1+3" hybrid, We marked numbers with bold where improvements come with $\alpha < 0.05$.

The best performing model is by far the second guess model, but since this model allows for two guesses, for fairness we omit it from our final comparison. To note, this second guess model corrected 110 incorrectly classified questions from the standard recognizer hybrid model.

For hybrid models, the one with raw rule features has better macro precision, but the one with recognizer response features achieves better macro recall. They both improve micro precision, recall and f1 score, compared to single models. Upon closer look, we found that the recognizer response features are able to increase the performance of the n-gram features on 'info/*' and 'fact/which' type questions, and the n-gram features are able to increase the performance of the recognizer response features on 'fact/polar and 'named entity/polar' type questions. The two feature sets complement each other well. It is also interesting to see that supplementing the n-grams with the raw rule features has similar micro-f1 and lower macro-f1, compared with supplementing the n-grams with recognizer features. This suggests that the partial order in aggregating individual output from recognizers in the rule-based classifier is likely to contribute to an overall improvement among question classes.

	Macro (%)			**Micro (%)**
Method	**Precision**	**Recall**	**F1**	**P/R/F1**
1. n-grams	68.8	60.9	61.7	82.9
2. raw rules	72.0	60.0	63.1	79.2
3. RBC	**79.8**	**79.5**	**77.1**	**84.8**
1+2	80.0	66.8	68.2	87.0
1+3	78.0	70.5	72.2	**87.0**
1+3 Top-2	**87.5**	**82.2**	**83.9**	**94.0**

Table 2: Performances of different feature configurations on the test set. Metrics include the macro- and micro-averaged precision, recall and f1-score.

For detailed comparison, Table 3 shows the per-class f-measures of different models. It is noted that n-grams alone tend to totally miss scarcely populated question classes. On the other hand, rule-based classifier achieves more balanced performances. In fact, there is no total miss of a single question class. For semantic-only classification, overall performance improves due to better populated classes. The rule-based model continues giving better macro-recall (83.2%) than the n-grams model (79.0%), thanks to the rules targeting the infrequent classes.

5. Conclusion and Future Work

We described the creation and annotation of a question classification corpus on enterprise content, as well as the question classification system used by the Watson Sales Assistant pilot project. Our question classification scheme assigns question classes by integrating semantic and syntactic characteristics. Expert sourcing question collection and annotation was carried out within the enterprise. Enhanced inter-annotator agreement monitoring and disagreement resolving were performed throughout the process to guarantee the quality of question class annotations. Our classification guideline, question corpus and other supplemental materials can be downloaded at the following URL: https://ibm.biz/BdRnaH.

Models	appr/how	deg/how	fact/what	rel/how	loc/where	appr/polar	appr/what	diff/how	fact/polar	ne/polar	ne/what
1. n-grams	0.8	0.91	0.69	0	0.78	0	0.22	0.33	0.85	0.67	0.76
2. raw rules	0.76	0.95	0.75	0.67	0.74	0.2	0.62	0.33	0.77	0.56	0.77
3. RBC	0.86	0.95	0.82	0.67	0.77	0.34	0.8	0.73	0.64	0.62	0.84
1+2	0.79	0.95	0.81	0	0.86	0	0.62	0.33	0.85	0.61	0.83
1+3	0.8	0.95	0.83	0	0.75	0	0.71	0.33	0.85	0.7	0.85
1+3+ 2nd guess	1	0.95	0.91	0.33	0.93	0.77	0.71	0.29	0.93	0.92	0.94

Models	info/what	ne/which	def/what	deg/what	diff/what	ne/who	info/polar	reason/why	fact/which	info/where	time/when
1. n-grams	0	0.87	0.94	0.73	0.9	0.98	0.25	1	0	0.91	0.98
2. raw rules	0	0.43	0.86	0.78	0.85	0.97	0.22	0.84	0	0.88	0.94
3. RBC	0.6	0.82	0.9	0.76	0.85	0.99	0.73	0.95	0.52	0.84	0.98
1+2	0	0.93	0.94	0.81	0.93	0.97	0.25	1	0.62	0.92	0.98
1+3	0.44	0.89	0.94	0.83	0.86	0.98	0.75	0.89	0.57	0.89	0.97
1+3+ 2nd guess	0.44	0.97	0.98	0.84	0.92	0.99	0.89	0.89	0.89	0.97	0.98

Table 3: The per-class f-measures of different models on test set. Classes are in the form of semantic/syntactic classification, as defined in section 3.2. Abbreviations used in the table: approach (appr), relation (rel), location (loc), difference (diff), named entity (ne), definition (def), degree (deg).

The fact that questions are collected from SMEs may bias the corpus toward questions that already have answers and may not reflect the questions from all salespeople. We are expanding the question collection by opening the question collection UI to the rest of salespeople even if they do not have an answer. We are continuing to collect and annotate more questions and expect to release a larger data set with more populated question classes. We also monitored the ratio of under-populated classes to guide possible merging (current frequency threshold 10, adjustable).

Evaluation with multiple rule-based and statistical classifiers showed that automated approach for classifying enterprise questions can achieve promising results and that hand crafted rules integrating syntactic, semantic and morphologic features do help. Including second guess improved the performance of the system significantly. When the top five predicted classes were included in the model, the performance increased to near perfection. This indicates that a feed-forward neural network or a deep learning architecture may perform well without a second guess model. We intend to evaluate these types of models in the future.

6. Acknowledgments

We would like to acknowledge the team of research scientists, engineers, and designers at IBM for their work to influence and develop extensions of Watson and the DeepQA architecture for enterprise use. It is this team who has made it possible for the work described in this paper. From IBM, Daniel Rucci, Alexander Rock, Meggan Todd, Desmond Wang, Dave Newbold, Edward Geraghty, Marita Beck, Beth Shiffman, Rob Johnson, Francoise Legoues.

7. References

Blunsom, P., Kocik, K. and Curran, J.R., (2006). Question classification with log-linear models. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 615–616.

Fellbaum, C., (1998). *WordNet: An Electronic Lexical Database*, Mit Press.

Fleiss, J.L., (1981). *Statistical Methods for Rates and Proportions*, Wiley.

Garcia-Fernandez, A., Rosset, S. and Vilnat, A., (2010). MACAQ: A Multi Annotated Corpus to Study how we Adapt Answers to Various Questions. In *LREC*. pp. 3559–3565.

Hacioglu, K. and Ward, W., (2003). Question classification with support vector machines and error correcting codes. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2*. Association for Computational Linguistics, pp. 28–30.

Hermjakob, U., (2001). Parsing and question classification for question answering. In *Proceedings of the workshop on Open-domain question answering-Volume 12*. Association for Computational Linguistics, pp. 1–6.

Hovy, E., Gerber, L., Hermjakob, U., Lin, C.-Y. and Ravichandran, D., (2001). Toward semantics-based answer pinpointing. In *Proceedings of the first international conference on Human language* *technology research*. Association for Computational Linguistics, pp. 1–7.

Hovy, E., Hermjakob, U. and Ravichandran, D., (2002). A question/answer typology with surface text patterns. In *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., pp. 247–251.

Li, X. and Roth, D., (2002). Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, pp. 1–7.

McCord, M.C., (2010). Using Slot Grammar. *IBM Technical Report*.

Moschitti, A., Chu-Carroll, J., Patwardhan, S., Fan, J. and Riccardi, G., (2011). Using syntactic and semantic structural kernels for classifying definition questions in Jeopardy! In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 712–724.

Moschitti, A., Quarteroni, S., Basili, R. and Manandhar, S., (2007). Exploiting syntactic and shallow semantic kernels for question answer classification. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*. pp. 776–783.

Murdock, J. and Welty, C., (2006). Obtaining Formal Knowledge from Informal Text Analysis. *IBM Technical Report*.

Ng, A.Y., (2004). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*. ACM, pp. 78–85.

Nguyen, M.L., Nguyen, T.T. and Shimazu, A., (2007). Subtree mining for question classification problem. In *Proceedings of the 20th international joint conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc., pp. 1695–1700.

Noreen, E.W., (1989). *Computer-intensive methods for testing hypotheses: an introduction*, Wiley.

Pinto, D. et al., (2002). QuASM: a system for question answering using semi-structured data. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM, pp. 46–55.

Ray, S.K., Singh, S. and Joshi, B., (2010). A semantic approach for question classification using WordNet and Wikipedia. *Pattern Recognition Letters*, 31, pp.1935–1943.

Solorio, T., Pérez-Coutino, M., Montes-y-Gómez, M., Villasenor-Pineda, L. and López-López, A., (2004). A language independent method for question classification. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, pp. 1374–1380.

Suchanek, F.M., Kasneci, G. and Weikum, G., (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*. ACM, pp. 697–706.

Voorhees, E.M., (2002). Overview of the TREC 2001 question answering track. *NIST special publication*, pp.42–51.

Voorhees, E.M., (1999). The TREC-8 question answering track report. In *Proceedings of TREC*. pp. 77–82.

Voorhees, E.M. and Tice, D., (2000). Overview of the TREC-9 question answering track. In *Proceedings of TREC*. pp. 71–80.

Zhang, D. and Lee, W.S., (2003). Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, pp. 26–32.