Automatic Identification of Rhetorical Questions

Shohini Bhattasali Dept. of Linguistics Cornell University Ithaca, NY, USA

Jeremy Cytryn Dept. of Computer Science Cornell University Ithaca, NY, USA

Elana Feldman Dept. of Linguistics Cornell University Ithaca, NY, USA

Joonsuk Park Dept. of Computer Science Cornell University Ithaca, NY, USA

{sb2295, jmc677, eaf82}@cornell.edu jpark@cs.cornell.edu

Abstract

A question may be asked not only to elicit information, but also to make a statement. Questions serving the latter purpose, called rhetorical questions, are often lexically and syntactically indistinguishable from other types of questions. Still, it is desirable to be able to identify rhetorical questions, as it is relevant for many NLP tasks, including information extraction and text summarization. In this paper, we explore the largely understudied problem of rhetorical question identification. Specifically, we present a simple n-gram based language model to classify rhetorical questions in the Switchboard Dialogue Act Corpus. We find that a special treatment of rhetorical questions which incorporates contextual information achieves the highest performance.

1 Introduction

Rhetorical questions frequently appear in everyday conversations. A rhetorical question is functionally different from other types of questions in that it is expressing a statement, rather than seeking information. Thus, rhetorical questions must be identified to fully capture the meaning of an utterance. This is not an easy task; despite their drastic functional differences, rhetorical questions are formulated like regular questions.

Bhatt (1998) states that in principle, a given question can be interpreted as either an information seeking question or as a rhetorical question and that intonation can be used to identify the interpretation intended by the speaker. For instance, consider the following example:

(1) Did I tell you that writing a dissertation was easy?

Just from reading the text, it is difficult to tell whether the speaker is asking an informational question or whether they are implying that they did not say that writing a dissertation was easy.

However, according to our observation, which forms the basis of this work, there are two cases in which rhetorical questions can be identified solely based on the text. Firstly, certain linguistic cues make a question obviously rhetorical, which can be seen in examples (2) and (3)[1]. Secondly, the context, or neighboring utterances, often reveal the rhetorical nature of the question, as we can see in example (4).

(2) Who ever lifted a finger to help George?

(3) After all, who has any time during the exam period?

(4) Who likes winter? It is always cold and windy and gray and everyone feels miserable all the time.

There has been substantial work in the area of classifying dialog acts, within which rhetorical questions fall. To our knowledge, prior work on dialog act tagging has largely ignored rhetorical questions, and there has not been any previous work specifically addressing rhetorical question identification. Nevertheless, classification of rhetorical questions is crucial and has numerous potential applications, including question-answering, document summarization, author identification, and opinion extraction.

We provide an overview of related work in Section 2, discuss linguistic characteristics of rhetorical questions in Section 3, describe the experimental setup in Section 4, and present and analyze the experiment results in Section 5. We find that, while the majority of the classification relies on features extracted from the question itself, adding

[1] See Section 3 for more details.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 743–749, Beijing, China, July 26-31, 2015.

in n-gram features from the context improves the performance. An F_1-score of 53.71% is achieved by adding features extracted from the preceding and subsequent utterances, which is about a 10% improvement from a baseline classifier using only the features from the question itself.

2 Related work

Jurafsky et al. (1997a) and Reithinger and Klesen (1997) used n-gram language modeling on the Switchboard and Verbmobil corpora respectively to classify dialog acts. Grau et al. (2004) uses a Bayesian approach with n-grams to categorize dialog acts. We also employ a similar language model to achieve our results.

Samuel et al. (1999) used transformation-based learning on the Verbmobil corpus over a number of utterance features such as utterance length, speaker turn, and the dialog act tags of adjacent utterances. Stolcke et al. (2000) utilized Hidden Markov Models on the Switchboard corpus and used word order within utterances and the order of dialog acts over utterances. Zechner (2002) worked on automatic summarization of open-domain spoken dialogues i.e., important pieces of information are found in the back and forth of a dialogue that is absent in a written piece.

Webb et al. (2005) used intra-utterance features in the Switchboard corpus and calculated n-grams for each utterance of all dialogue acts. For each n-gram, they computed the maximal predictivity i.e., its highest predictivity value within any dialogue act category. We utilized a similar metric for n-gram selection.

Verbree et al. (2006) constructed their baseline for three different corpora using the performance of the LIT set, as proposed by Samuel (2000). In this approach, they also chose to use a compressed feature set for n-grams and POS n-grams. We chose similar feature sets to classify rhetorical questions.

Our work extends these approaches to dialog act classification by exploring additional features which are specific to rhetorical question identification, such as context n-grams.

3 Features for Identifying Rhetorical Questions

In order to correctly classify rhetorical questions, we theorize that the choice of words in the question itself may be an important indicator of speaker intent. To capture intent in the words themselves, it makes sense to consider a common unigram, while a bigram model will likely capture short phrasal cues. For instance, we might expect the existence of n-grams such as *well* or *you know* to be highly predictive features of the rhetorical nature of the question.

Additionally, some linguistic cues are helpful in identifying rhetorical questions. Strong negative polarity items (NPIs), also referred to as emphatic or even-NPIs in the literature, are considered definitive markers. Some examples are *budge an inch*, *in years*, *give a damn*, *bat an eye*, and *lift a finger* (Giannakidou 1999, van Rooy 2003). Gresillon (1980) notes that a question containing a modal auxiliary, such as *could* or *would*, together with negation tends to be rhetorical. Certain expressions such as *yet* and *after all* can only appear in rhetorical questions (Sadock 1971, Sadock 1974). Again, using common n-grams as features should partially capture the above cues because n-gram segments of strong NPIs should occur more frequently.

We also wanted to incorporate common grammatical sequences found in rhetorical questions. To that end, we can consider part of speech (POS) n-grams to capture common grammatical relations which are predictive.

Similarly, for rhetorical questions, we expect context to be highly predictive for correct classification. For instance, the existence of a question mark in the subsequent utterance when spoken by the questioner, will likely be a weak positive cue, since the speaker may not have been expecting a response. However, the existence of a question mark by a different speaker may not be indicative. This suggests a need to decompose the context-based feature space by speaker. Similarly, phrases uttered prior to the question will likely give rise to a different set of predictive n-grams.

Using these observations, we decided to implement a simple n-gram model incorporating contextual cues to identify rhetorical questions. Specifically, we used unigrams, bigrams, POS bigrams, and POS trigrams of a question and its immediately preceding and following context as feature sets. Based on preliminary results, we did not use trigrams or POS unigrams. POS tags did not capture sufficient contextual information and trigrams were not implemented since the utterances in our dataset were too small to fully utilize them.

Also, to capture the contextual information, we

distinguish three distinct categories - questions, utterances immediately preceding questions, and utterances immediately following questions. In order to capture the effect of a feature if it is used by the same speaker versus a different speaker, we divided the feature space contextual utterances into four disjoint groups: *precedent-same-speaker*, *precedent-different-speaker*, *subsequent-same-speaker*, and *subsequent-different-speaker*. Features in each group are all considered independently.

4 Experimental Setup

4.1 Data

For the experiments, we used the Switchboard Dialog Act Corpus (Godfrey et al. 1992; Jurafsky et al. 1997b), which contains labeled utterances from phone conversations between different pairs of people. We preprocessed the data to contain only the utterances marked as questions (rhetorical or otherwise), as well as the utterances immediately preceding and following the questions. Additionally, connectives like *and* and *but* were marked as *t_con*, the end of conversation was marked as *t_empty*, and laughter was marked as *t_laugh*.

After filtering down to questions, we split the data into 5960 questions in the training set and 2555 questions in the test set. We find the dataset to be highly skewed with only $\frac{128}{2555}$ or 5% of the test instances labeled as rhetorical. Because of this, a classifier that naively labels all questions as non-rhetorical would achieve a 94.99% accuracy. Thus, we chose precision, recall and F_1-measure as more appropriate metrics of our classifier performance. We should note also that our results assume a high level of consistency of the hand annotations from the original taggging of the Switchboard Corpus. However, based on our observation and the strict guidelines followed by annotators as mentioned in Jurafsky et al. (1997a), we are reasonably confident in the reliability of the rhetorical labels.

4.2 Learning Algorithm

We experimented with both Naive Bayes and a Support Vector Machine (SVM) classifiers. Our Naive Bayes classifier was smoothed with an add-alpha Laplacian kernel, where alpha was selected via cross-validation. For our SVM, to account for the highly skewed nature of our dataset, we set the cost-factor based on the ratio of positive (rhetorical) to negative (non-rhetorical) questions in our training set as in Morik et al. (1999). We tuned the trade-off between margin and training error via cross validation over the training set.

In early experiments, Naive Bayes performed comparably to or outperformed SVM because the dimensionality of the feature space was relatively low. However, we found that SVM performed more robustly over the large range and dimensionality of features we employed in the later experiments. Thus, we conducted the main experiments using SVMLite (Joachims 1999).

As the number of parameters is linear in the number of feature sets, an exhaustive search through the space would be intractable. So as to make this feasible, we employ a greedy approach to model selection. We make a naive assumption that parameters of feature sets are independent or codependent on up to one other feature set in the same group. Each pair of codependent feature sets is considered alone while holding other feature sets fixed. Classifier parameters are also assumed to be independent for tuning purposes.

In order to optimize search time without sampling the parameter space too coarsely, we employed an adaptive refinement variant to a traditional grid search. First, we discretely sampled the Cartesian product of dependent parameters sampled at regular geometric or arithmetic intervals between a user-specified minimum and maximum. We then updated minimum and maximum values to center around the highest scoring sample and recursed on the search with the newly downsized span for a fixed recursion depth d. In practice, we choose $k = 4$ and $d = 3$.

4.3 Features

Unigrams, bigrams, POS bigrams, and POS trigrams were extracted from the questions and neighboring utterances as features, based on the analysis in Section 3. Then, feature selection was performed as follows.

For all features sets, we considered both unigram and bigram features. All unigrams and bigrams in the training data are considered as potential candidates for features. For each feature set above, we estimated the maximal predictivity over both rhetorical and non-rhetorical classes, corresponding to using the MLE of $P(c|n)$, where n denotes the n-gram and c is the class. We used these estimates as a score and select the j n-grams

with the highest score for each *n* over each group, regardless of class, where *j* was selected via 4-fold cross validation.

Each feature was then encoded as a simple occurrence count within its respective group for a given exchange. The highest scoring unigrams and bigrams are as follows: "you", "do", "what", "to", "t_con", "do you", "you know", "going to", "you have", and "well ,".

POS features were computed by running a POS tagger on all exchanges and and then picking the *j*-best n-grams as described above. For our experiments, we used the maximum entropy treebank POS tagger from the NLTK package (Bird et al. 2009) to compute POS bigrams and trigrams.

Lastly, in order to assess the relative value of question-based and context-based features, we designed the following seven feature sets:

- Question (*baseline*)
- Precedent
- Subsequent
- Question + Precedent
- Question + Subsequent
- Precedent + Subsequent
- Question + Precedent + Subsequent

The question-only feature set serves as our baseline without considering context, whereas the other feature sets serve to test the power of the preceding and following context alone and when paired with features from the question itself.

Feature set	Acc	Pre	Rec	F1	Error 95%
Question	92.41	35.00	60.16	44.25	7.59 ±1.02
Precedent	85.64	12.30	30.47	17.53	14.36 ±1.36
Subsequent	78.98	13.68	60.16	22.29	21.02 ±1.58
Question + Precedent	93.82	41.94	60.94	49.68	6.18 ±0.93
Question + Subsequent	93.27	39.52	**64.84**	49.11	6.73 ±0.97
Precedent + Subsequent	84.93	19.62	**64.84**	30.14	15.07 ±1.38
Question + Precedent + Subsequent	**94.87**	**49.03**	59.38	**53.71**	**5.13± 0.86**

Table 1: Experimental results (%)

AC	PC	Utterance
+	+	X: 'i mean, why not.'
	-	X: 'what are you telling that student?'
-	+	X: '*t_laugh* why don't we do that?'
	-	X: 'who, was in that.'

Table 2: Classification without Context Features (AC: Actual Class, P: Predicted Class. X denotes the speaker)

AC	PC	Utterances
+	+	X: '*t_con* you give them an f on something that doesn't seem that bad to me.' **X: 'what are you telling that student?'** X: 'you're telling them that, hey, you might as well forget it, you know.'
	-	X: 'get homework done,' **X: '*t_con* you know, where do you find the time'.** Y:'well, in the first place it's not your homework,'
-	+	X: 'ha, ha, lots of luck.' **X: 'is she spayed.'** Y: 'yeah'.
	-	Y: '*t_con* it says when the conversation is over just say your good-byes and hang up.' **X: '*t_laugh* why don't we do that?** Y: 'i, guess so.'

Table 3: Classification with Context Features (AC: Actual Class, PC: Predicted Class. X and Y denote the speakers)

5 Results and Analysis

Table 1 shows the performance of the feature sets cross-valided and trained on 5960 questions (with context) in the Switchboard corpus and tested on the 2555 remaining questions.

Our results largely reflect our intuition on the expected utility of our various feature sets. Features in the question group prove by far the most useful single source, while features within the subsequent prove to be more useful than features in the precedent. Somewhat surprisingly however, an F_1-score of 30.14% is achieved by training on contextual features alone while ignoring any cues from the question itself, suggesting the power of context in identifying a question as rhetorical. Additionally, one of the highest scoring bigrams is *you know*, matching our earlier intuitions.

Some examples of the success and failings of our system can be found in Table 2 and 3. For instance, in our question-only feature space, the phrase *what are you telling that student?* was incorrectly classified as non-rhetorical. When the contextual features were added in, the classifier correctly identified it as rhetorical as we might expect. Failure cases of our simple language model based system can be seen for instance in the false positive question *is she spayed* which is inter-

preted as rhetorical, likely due to the unigram *yeah* in the response.

Overall, we achieve our best results when including both precedent and subsequent context along with the question in our feature space. Thus, our results suggest that incorporating contextual cues from both directly before and after the question itself outperforms classifiers trained on a naive question-only feature space.

5.1 Feature Dimensionality

After model selection via cross validation, our total feature space dimensionality varies between 2914 for the *precedent* only feature set and 16615 for the *question + subsequent* feature set. Distinct n-gram and POS n-gram features are considered for each of same speaker and different speaker for precedents and subsequents so as to capture the distinction between the two. Examining the relative number of features selected for these subfeature sets also gives a rough idea of the strength of the various cues. For instance, same speaker feature dimensionality tended to be much lower than different speaker feature dimensionality, suggesting that considering context uttered by the respondent is a better cue as to whether the question is rhetorical. Additionally, unigrams and bigrams tend to be more useful features than POS n-grams for the task of rhetorical question identification, or at least considering the less common POS n-grams is not as predictive.

5.2 Evenly Split Distribution

As the highly skewed nature of our data does not allow us to get a good estimate of error rate, we also tested our feature sets on a subsection of the dataset with a 50-50 split between rhetorical and non-rhetorical questions to get a better sense of the accuracy of our classifier. The results can be seen in Table 4. Our classifier achieves an accuracy of 81% when trained on the questions alone and 84% when integrating precedent and subsequent context. Due to the reduced size of the evenly split dataset, performing a McNemar's test with Edwards' correction (Edwards 1948) does not allow us to reject the null hypothesis that the two experiments do not derive from the same distribution with 95% confidence ($\chi^2 = 1.49$ giving a 2-tailed p value of 0.22). However, over the whole skewed dataset, we find $\chi^2 = 30.74$ giving a 2-tailed $p < 0.00001$ so we have reason to believe that with a larger evenly-split dataset integrating context-based features provides a quantifiable advantage.

Feature set	Acc	Pre	Rec	F1	Error 95%
Question	81.25	82.71	78.01	80.29	0.19 $\pm$0.05
Question + Precedent + Subsequent	84.38	88.71	78.01	83.02	0.16 $\pm$0.04

Table 4: Experimental results (%) on evenly distributed data (training set size: 670 & test set size: 288)

6 Conclusions

In this paper, we tackle the largely understudied problem of rhetorical question identification. While the majority of the classification relies on features extracted from the question itself, adding in n-gram features from the context improves the performance. We achieve a 53.71% F_1-score by adding features extracted from the preceding and the subsequent utterances, which is about a 10% improvement from a baseline classifier using only the features from the question itself.

For future work, we would like to employ more complicated features like the sentiment of the context, and dictionary features based on an NPI lexicon. Also, if available, prosodic information like focus, pauses, and intonation may be useful.

7 Acknowledgements

We thank Mary Moroney and Andrea Hummel for helping us identify linguistic characteristics of rhetorical questions and the anonymous reviewers for their thoughtful feedback.

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *ICASSP (1)*, pages 1061–1064.

Rajesh Bhatt. 1998. Argument-adjunct asymmetries in rhetorical questions. In *NELS 29*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Allen L. Edwards. 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. In *Psychometrika*, 13(3):185–187.

Anastasia Giannakidou. 1999. Affective dependencies In *Linguistics and Philosophy*, 22(4): 367–421. Springer

John J. Godfrey, Edward C. Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520 vol.1.

Sergio Grau, Emilio Sanchis, María José Castro, David Vilar. 2004. Dialogue act classification using a Bayesian approach In *9th Conference Speech and Computer*.

Almuth Gresillon. 1980. Zum linguistischen Status rhetorischer Fragen In*Zeitschrift für germanistische Linguistik*, 8(3): 273–289.

Chung-Hye Han. 1998. Deriving the interpretation of rhetorical questions. In *Proceedings of West Coast Conference in Formal Linguistics*, volume 16, pages 237–253. Citeseer.

T. Joachims. 1999. Making large-scale svm learning practical. Advances in kernel methods-support vector learning.

Dan Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, Carol V. Ess-Dykema, et al. 1997a. Automatic detection of discourse structure for speech recognition and understanding. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 88–95. IEEE.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997b. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report Draft 13, University of Colorado, Institute of Cognitive Science.

Simon Keizer, Anton Nijholt, et al. 2002. Dialogue act recognition with bayesian networks for dutch dialogues. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2*, pages 88–94. Association for Computational Linguistics.

Katharina Morik, Peter Brockhausen, and T. Joachims. 1999. Combining statistical learning with a knowledge-based approach: a case study in intensive care monitoring. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund.

Norbert Reithinger and Martin Klesen. 1997. Dialogue act classification using language models. In *EuroSpeech*. Citeseer.

Jerrold M. Saddock. 1971. Queclaratives In *Seventh Regional Meeting of the Chicago Linguistic Society*, 7: 223–232.

Jerrold M. Saddock. 1974. Toward a linguistic theory of speech acts Academic Press New York

Ken Samuel, Sandra Carberry, and K. Vijay-Shanker. 1999. Automatically selecting useful phrases for dialogue act tagging. *arXiv preprint cs/9906016*.

Ken B. Samuel. 2000. *Discourse learning: an investigation of dialogue act tagging using transformation-based learning*. University of Delaware.

Elizabeth Shriberg, Andreas Stolcke, Dan Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Dan Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Robert van Rooy. 2003. Negative polarity items in questions: Strength as relevance In *Journal of Semantics*, 20(3): 239–273. Oxford University Press.

Anand Venkataraman, Andreas Stolcke, and Elizabeth Shriberg. 2002. Automatic dialog act labeling with minimal supervision. In *9th Australian International Conference on Speech Science and Technology, SST 2002*.

Daan Verbree, Rutger Rienks, and Dirk Heylen. 2006. Dialogue-act tagging using smart feature selection; results on multiple corpora. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 70–73. IEEE.

Volker Warnke, Ralf Kompe, Heinrich Niemann, and Elmar Nöth. 1997. Integrated dialog act segmentation and classification using prosodic features and language models. In *EUROSPEECH*.

Nick Webb, Mark Hepple, and Yorik Wilks. 2005. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*. Citeseer.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.

Matthias Zimmerman, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. 2005. A* based joint segentation and classification of dialog acts in multiparty meetings. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 215–219. IEEE.