Comparing Approaches for Automatic Question Identification

Angel Samsuddin Maredia, Kara Schechtman, Sarah Ita Levitan, Julia Hirschberg

Department of Computer Science, Columbia University, USA

asm2221@columbia.edu, kws2121@columbia.edu,
sarahita@cs.columbia.edu, julia@cs.columbia.edu

Abstract

Collecting spontaneous speech corpora that are open-ended, yet topically constrained, is increasingly popular for research in spoken dialogue systems and speaker state, inter alia. Typically, these corpora are labeled by human annotators, either in the lab or through crowdsourcing; however, this is cumbersome and time-consuming for large corpora. We present four different approaches to automatically tagging a corpus when general topics of the conversations are known. We develop these approaches on the Columbia X-Cultural Deception corpus and find accuracy that significantly exceeds the baseline. Finally, we conduct a cross-corpus evaluation by testing the best performing approach on the Columbia/SRI/Colorado corpus.

1 Introduction

Corpora of spontaneous speech are often collected through interviews or by otherwise providing subjects with question prompts. Such corpora are semi-structured; they are constrained by the prompts used, but the elicited speech is open-ended in vocabulary and structure. It is often desirable to segment these corpora into their underlying topics based on the questions asked. This is typically done manually by annotators in the lab or via crowd-sourcing. However, such annotation is impractical and time-consuming for large corpora.

In this paper we describe a set of experiments aimed at automatically tagging a large corpus with topic labels. We tag the Columbia X-Cultural Deception (CXD) corpus, a large-scale (120-hour) corpus of deceptive and non-deceptive dialogues collected using a semi-structured interview paradigm. Participants took turns interviewing each other using a fixed set of biographical interview questions[1], but the questions were asked in individual variants, in any order, and interviewers often asked follow-up questions. For example, the question, "Are your parents divorced?" could be produced as "Are your mom and dad still together?" These questions are semantically similar, but differ lexically, presenting the challenge of topically tagging a corpus based on semantic similarity. The question, "Have you ever broken a bone?" could be followed by another, "How did you break your bone?" This illustrates the challenge of distinguishing between phrases that are lexically similar, but differ semantically. These two examples highlight problems faced when trying to automatically annotate a corpus for responses to a given set of questions.

With such a large corpus, it is not practical to manually annotate topic boundaries. So, to compare question responses from multiple subjects, we identify conversational turns in the corpus that correspond to the original interview questions. We compare four approaches to question identification: (1) a baseline approach that identifies questions using strict string matches, (2) the ROUGE metric which is based on n-gram comparisons, (3) cosine similarity between word embedding representations and (4) cosine similarity between document embeddings. We include experiments with varying thresholds for approaches (2), (3), and (4) to highlight the trade-off between precision and recall for these approaches. Finally, we test our best approach using word embeddings on another corpus, the Columbia/SRI/Colorado (CSC) corpus (Hirschberg et al., 2005), collected with a similar interview paradigm but different questions, in order to evaluate the utility of this method in another

[1]The interview questions can be found here: http://tinyurl.com/lzfa8zl

domain.

This work draws upon the body of research on short-text semantic similarity (e.g. (Mihalcea et al., 2006; Kenter and de Rijke, 2015; Oliva et al., 2011)). It is also related to work on topic segmentation (e.g. (Cardoso et al., 2013; Dias et al., 2007)), however here we focus on matching conversational turns to a fixed set of possible topics. While this work is done in support of our ongoing work on deception detection using speech and text-based features, we believe that our approach could be applied to other spontaneous transcribed speech or text corpora which were collected with some constraints on topics.

2 Corpus

The Columbia X-Cultural Deception (CXD) Corpus (Levitan et al., 2015) is a collection of within-subject deceptive and non-deceptive speech from native speakers of Standard American English (SAE) and Mandarin Chinese (MC), all speaking in English. The corpus contains dialogues between 340 subjects. A variation of a fake resume paradigm was used to collect the data. Previously unacquainted pairs of subjects played a "lying game" with each other. Each subject filled out a 24-item biographical questionnaire and were instructed to create false answers for a random half of the questions. The lying game was recorded in a sound booth. For the first half of the game, one subject assumed the role of the interviewer, while the other answered the biographical questions, lying for half and telling the truth for the other; questions chosen in each category were balanced across the corpus. For the second half of the game, the subjects roles were reversed, and the interviewer became the interviewee. During the game, the interviewer was allowed to ask the 24 questions in any order s/he chose; the interviewer was also encouraged to ask follow-up questions to aid them in determining the truth of the interviewees answers. The entire corpus was orthographically transcribed using the Amazon Mechanical Turk (AMT)[2] crowd-sourcing platform, and transcripts were forced-aligned with the audio recordings. The speech was then automatically segmented into *inter-pausal units* (IPUs), defined as pause-free segments of speech separated by a minimum pause length of 50 ms. The speech was also segmented into turn units, where a turn is defined as a maximal sequence of IPUs from a single speaker without any interlocutor speech that is not a *backchannel* (a simple acknowledgment that is not an attempt to take the turn). For this work, we compiled 40 interviewer sessions (about 20% of the corpus) and hand-annotated the turns for all of these sessions, giving us a total of 5308 turns. Out of these turns, 923 were interviewer questions that corresponded to the list of the original biographical questions, which we labeled with the question number. Below we describe the different approaches and then discuss results in Section 4 with a comparison of performance in Table 1.

[2]https://www.mturk.com/mturk/

3 Question Identification Approaches

3.1 String-matching Baseline

As a baseline for matching the 24 questions interviewers were instructed to ask with interviewer turns, we performed a simple two-pass question matching procedure for exact string matches between written questions and the transcripts. In the first pass, we searched for exact matches of strings with punctuation and spacing removed. With the remaining unmatched questions, we then performed another round of matching, with the transcript lemmatized and with filler words removed, to identify very close though not exact matches.

3.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)(Lin, 2004) is a package designed to evaluate computer-generated summaries against a human-written baseline using a simple n-gram comparison to find precision, recall, and f-score for each machine-human summary comparison. Using ROUGE, we evaluated matches for questions which had not been detected by the baseline. We created a ROUGE task for each unmatched question. For each task, the original question was used as the reference text. We then tested each interviewer turn in the conversation against the reference, using bi-gram matching. We thus matched the turn receiving the highest similarity score to the reference text to that question, testing this method at a variety of similarity thresholds.

3.3 Word Embeddings

The previous two methods identify questions using lexical similarity. In the next two approaches we explored semantic similarity. We began by obtaining a vector representation for each of the

24 questions. We use a pre-trained Word2vec model on the Google News dataset[3] with over three million words and phrases to obtain word embeddings. The primary benefit of a Word2vec model is that it clusters semantically similar words and phrases together: for example, "Golden Gate Bridge" and "San Francisco" have very low cosine distance between each other in this model. Therefore, semantically similar words were likely to be represented as vectors with high cosine similarity.

To obtain a vector representation for each question as a whole, we found the vector representation for each word using Word2vec. We then took a weighted average of all of the word vectors in the question where words that directly contributed to the topic of the turn such as "relationship" or "mom" were weighed more than words that, if removed, did not affect the topic of the turn such as "have" or "really." This produced a final vector representation of the entire question. We exclude stop words from this vector average. Following the same approach, we obtained vector representations for each interviewer turn. We then calculated the cosine similarities between a turn and each question and found the question that had the highest cosine similarity to the turn vector. We compared the cosine similarity of the turn and the question to the cosine similarity of any previous identified matches. If the newly calculated cosine similarity was higher, then the current turn was deemed the best match so far to the question, otherwise we repeated this comparison with the question that had the second highest cosine similarity to the turn. At the end of each particular interviewer session, we had a mapping of each turn to a question if a match was detected, otherwise the turn was marked as not being a question.

3.4 Document Embeddings

We also explored the use of document embeddings for this task. We began by finding a vector representation for each of the 24 questions. We used a Doc2vec model pre-trained on Wikipedia text[4]. Recall that, in our paradigm, questions could be asked in individual variants, in any order, and along with follow-up questions. The primary benefit of a Doc2Vec model is that it allows for unsupervised learning of larger blocks of text. Therefore, we hypothesized that Doc2Vec would return word vectors that also depended on contextual usage as well as semantic similarity. We then calculated the vector averages for each turn and produced turn-to-question mappings as explained in the word embeddings approach above.

[3]The model can be found here: https://code.google.com/archive/p/word2vec/

[4]The model can be found here: https://github.com/jhlau/doc2vec

4 Results

Table 1 shows the accuracy, precision, recall, and f1-score of each of the four approaches outlined above, evaluated on our hand-labeled subset of interviews. We see that the word embeddings method achieved the highest accuracy, recall, and f1-score of all the methods developed and tested, whereas the ROUGE approach obtained the highest precision. With the word embeddings approach, most correctly identified turns share one or more meaningful words with the corresponding original question and are often syntactically very similar. This approach, however, is able to make ambiguous matches as well. For example, an interviewer turn said, "wow you broke you broke your hand when you were in elementary school wow i yeah i get so student hate to do homework so have you ever tweet tweeted." This turn shares meaningful words with many other questions, but this approach correctly identified it as matching the question Have you ever tweeted? The word embeddings approach could also make difficult semantic matches. Many interviewers asked, "Are your mom and dad still together?" instead of "Are your parents divorced?" Even though there are few lexically common meaningful words between these two phrases, this approach correctly mapped these questions to each other because of their semantic similarity. One of the main causes of error for this method is that follow-up questions were sometimes mis-identified as original questions. For example, "How do you like your major?" could be mapped to the original question, "If you attended college, what was your major?" even though the question the interviewer asked was a follow-up question.

We also analyzed the accuracy of the methodologies using varying thresholds. For word embeddings and document embeddings, the threshold is determined by cosine similarity of a turn and question. For ROUGE, the threshold is the f1-score. For each approach, We compiled a set of turns from the CXD corpus that had the lowest cosine similarity to the question each turn was

Approach	Accuracy	Precision	Recall	F1-Score
Baseline (Rule-based)	39.0	72.0	42.0	53.1
ROUGE	74.0	**93.0**	78.0	84.8
Word Embeddings	**91.4**	92.1	**99.1**	**95.5**
Document Embeddings	88.6	90.0	98.2	93.9

Table 1: Accuracy, precision, and recall of each approach, evaluated on hand annotated turns

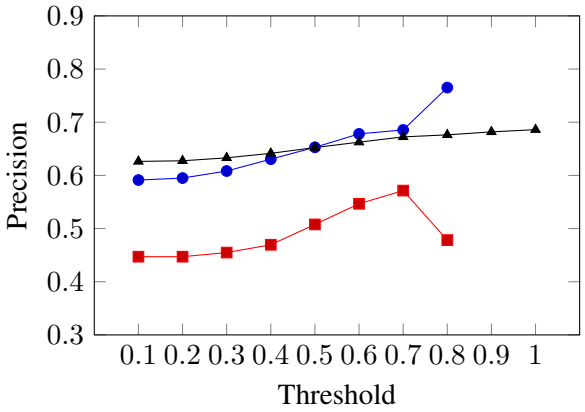

Figure 1: Accuracy of each approach determined by threshold. Filled in dots are word embeddings. Squares represent document embeddings. Triangles represent ROUGE.

matched with. We capped the threshold at 0.82. Figure 1 shows that, as we increase the threshold, generally, the accuracy of the question matching for all approaches is higher. This, intuitively, makes sense because, as we increase the threshold, we are selecting turns that have higher similarity to their matched question. Although this results in lower recall, it can be used in cases where high precision is needed for annotations.

4.1 Cross-corpus Evaluation

To further evaluate our best-performing approach, we applied the word embeddings method to another corpus collected using a similar interview paradigm, the Columbia SRI Colorado (CSC) corpus. To test word embeddings on this corpus, we compiled 31 interviewer sessions that were already hand annotated, giving us a total of 6395 turns. The (single) interviewer involving in collecting this corpus always began with a list of four standard biographical questions, thus reducing the number of turns that contained an interviewer-generated question to 114. Following the word embeddings method described above, we obtained an accuracy of 99.8%, precision of 91.2%, recall of 100%, and F1-score of 95.3 on the CSC corpus.

The incorrectly identified questions were largely because the interviewer did not ask all four biographical questions in every session, while the word embeddings approach assumes that all questions were asked and therefore, matches some turn to the original question even though the interviewer did not ask it. The higher accuracy obtained on the CSC corpus is probably due to the fact that the interviews were all conducted by a single interviewer, so the questions were asked with greater consistency. In addition, all subjects were native speakers of Standard American English, while half the participants in the CXD corpus were native speakers of Mandarin Chinese.

5 Conclusion

Corpora consisting of spontaneous speech that is open-ended, yet topically constrained, is more commonplace, as researchers seek spontaneous speech with some similarity of topic across subjects. Traditionally, such corpora are hand annotated for topic segments to serve as training material. However, on large corpora such as the CXD corpus, this can be cumbersome and time-consuming. In this paper, we have presented four approaches to automatically identifying question topics on the CXD corpus to discover which approach achieves the best results in automatically tagging corpora into question-defined topics. We found that the word embeddings approach was the best performing approach with an f1-score of 95.5%. We then applied the word embeddings approach to the CSC corpus to verify that this approach was useful for other corpora and also achieved very good results. We conclude that this automated, unsupervised approach to tagging corpora can be very useful in annotation and analysis for corpora collected using question prompts. For more exact annotations, this approach could also be used as an automated pre-processing stage to reduce human annotation efforts. In future, we would like to extend the embeddings approach to scale to less constrained tasks, evaluate it on additional corpora, and also more accurately tag corpora based on an ambiguous number of topics.

References

Paula CF Cardoso, Maite Taboada, and Thiago AS Pardo. 2013. Subtopics annotation in a corpus of news texts: steps towards automatic subtopic segmentation. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*.

Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes. 2007. Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation. In *AAAI*. volume 7, pages 1334–1340.

Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. 2005. Distinguishing deceptive from non-deceptive speech. In *Interspeech*. pages 1833–1836.

Tom Kenter and Maarten de Rijke. 2015. Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, pages 1411–1420.

Sarah I Levitan, Guzhen An, Mandi Wang, Gideon Mendels, Julia Hirschberg, Michelle Levine, and Andrew Rosenberg. 2015. Cross-cultural production and detection of deception from speech. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. ACM, pages 1–8.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*. Barcelona, Spain, volume 8.

Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*. volume 6, pages 775–780.

Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, and Ángel Iglesias. 2011. Symss: A syntax-based measure for short-text semantic similarity. *Data & Knowledge Engineering* 70(4):390–405.