

Off-Topic Detection in Conversational Telephone Speech

Robin Stewart and **Andrea Danyluk**

Department of Computer Science

Williams College

Williamstown, MA 01267

{06rssl_2, andrea}@cs.williams.edu

Yang Liu

Department of Computer Science

UT Dallas

Richardson, TX 75080

yangl@hlt.utdallas.edu

Abstract

In a context where information retrieval is extended to spoken “documents” including conversations, it will be important to provide users with the ability to seek informational content, rather than socially motivated small talk that appears in many conversational sources. In this paper we present a preliminary study aimed at automatically identifying “irrelevance” in the domain of telephone conversations. We apply a standard machine learning algorithm to build a classifier that detects off-topic sections with better-than-chance accuracy and that begins to provide insight into the relative importance of features for identifying utterances as on topic or not.

1 Introduction

There is a growing need to index, search, summarize and otherwise process the increasing amount of available broadcast news, broadcast conversations, meetings, class lectures, and telephone conversations. While it is clear that users have wide ranging goals in the context of information retrieval, we assume that some will seek only credible information about a specific topic and will not be interested in the socially-motivated utterances which appear throughout most conversational sources. For these users, a search for information about weather should not return conversations containing small talk such as “Nice weather we’ve been having.”

In this paper we investigate one approach for automatically identifying “irrelevance” in the domain of telephone conversations. Our initial data consist of conversations in which each utterance is labeled as being on topic or not. We apply inductive classifier learning algorithms to identify useful features and build classifiers to automatically label utterances.

We begin in Section 2 by hypothesizing features that might be useful for the identification of irrelevant regions, as indicated by research on the linguistics of conversational speech and, in particular, small talk. Next we present our data and discuss our annotation methodology. We follow this with a description of the complete set of features and machine learning algorithms investigated. Section 6 presents our results, including a comparison of the learned classifiers and an analysis of the relative utility of various features.

2 Linguistics of Conversational Speech

Cheepen (Cheepen, 1988) posits that speakers have two primary goals in conversation: **interactional** goals in which interpersonal motives such as social rank and trust are primary; and **transactional** goals which focus on communicating useful information or getting a job done. In a context where conversations are indexed and searched for information, we assume in this paper that users will be interested in the communicated information, rather than the way in which participants interact. Therefore, we assume that utterances with primarily transactional purposes will be most important, while interactional utterances can be ignored.

Greetings and partings are the most predictable

type of interactional speech. They consistently appear at the beginning and end of conversations and follow a fairly formulaic pattern of content (Laver, 1981). Thus we hypothesize that: *Utterances near the beginning or end of conversations are less likely to be relevant.*

Cheepen also defines **speech-in-action** regions to be segments of conversation that are related to the present physical world or the activity of chatting, e.g. “What lovely weather.” or “It is so nice to see you.” Since these regions mainly involve participants identifying their shared social situation, they are not likely to contain transactional content. Further, since speech-in-action segments are distinguished by their focus on the present, we hypothesize that: *Utterances with present tense verbs are less likely to be relevant.*

Finally, small talk that is not intended to demarcate social hierarchy tends to be abbreviated, e.g. “Nice day” (Laver, 1981). From this we hypothesize that: *Utterances lacking common helper words such as “it”, “there”, and forms of “to be” are less likely to be relevant.*

3 Related Work

Three areas of related work in natural language processing have been particularly informative for our research.

First, speech act theory states that with each utterance, a conversant is committing an action, such as questioning, critiquing, or stating a fact. This is quite similar to the notion of transactional and interactional goals. However, speech acts are generally focused on the lower level of breaking apart utterances and understanding their purpose, whereas we are concerned here with a coarser-grained notion of relevance. Work closer to ours is that of Bates et al. (Bates et al., 2005), who define **meeting acts** for recorded meetings. Of their tags, **commentary** is most similar to our notion of relevance.

Second, there has been research on *generating* small talk in order to establish rapport between an automatic system and human user (Bickmore and Cassell, 2000). Our work complements this by potentially detecting off-topic speech from the human user as an indication that the system should also respond with interactional language.

Label	Utterance
S	2: [LAUGH] Hi.
S	2: How nice to meet you.
S	1: It is nice to meet you too.
M	2: We have a wonderful topic.
M	1: Yeah.
M	1: It’s not too bad. [LAUGH]
T	2: Oh, I – I am one hundred percent in favor of, uh, computers in the classroom.
T	2: I think they’re a marvelous tool, educational tool.

Table 1: A conversation fragment with annotations: (S)mall Talk, (M)etaconversation, and On-(T)opic. The two speakers are identified as “1” and “2”.

Third, off-topic detection can be viewed as a segmentation of conversation into relevant and irrelevant parts. Thus our work has many similarities to topic segmentation systems, which incorporate cue words that indicate an abrupt change in topic (e.g. “so anyway...”), as well as long term variations in word occurrence statistics (Hearst, 1997; Reynar, 1999; Beeferman et al., 1999, e.g.). Our approach uses previous and subsequent sentences to approximate these ideas, but might benefit from a more explicitly segmentation-based strategy.

4 Data

In our work we use human-transcribed conversations from the Fisher data (LDC, 2004). In each conversation, participants have been given a topic to discuss for ten minutes. Despite this, participants often talk about subjects that are not at all related to the assigned topic. Therefore, a convenient way to define irrelevance in conversations in this domain is *segments which do not contribute to understanding the assigned topic*. This very natural definition makes the domain a good one for initial study; however, the idea can be readily extended to other domains. For example, broadcast debates, class lectures, and meetings usually have specific topics of discussion.

The primary transactional goal of participants in the telephone conversations is to discuss the assigned topic. Since this goal directly involves the act of discussion itself, it is not surprising that participants often talk about the current conversation or

the choice of topic. There are enough such segments that we assign them a special region type: **Metaconversation**. The purely irrelevant segments we call **Small Talk**, and the remaining segments are defined as **On-Topic**. We define utterances as segments of speech that are delineated by periods and/or speaker changes. An annotated excerpt is shown in Table 1.

For the experiments described in this paper, we selected 20 conversations: 4 from each of the topics “computers in education”, “bioterrorism”, “terrorism”, “pets”, and “censorship”. These topics were chosen randomly from the 40 topics in the Fisher corpus, with the constraint that we wanted to include topics that could be a part of normal small talk (such as “pets”) as well as topics which seem farther removed from small talk (such as “censorship”).

Our selected data set consists of slightly more than 5,000 utterances. We had 2-3 human annotators label the utterances in each conversation, choosing from the 3 labels Metaconversation, Small Talk, and On-Topic. On average, pairs of annotators agreed with each other on 86% of utterances. The main source of annotator disagreement was between Small Talk and On-Topic regions; in most cases this resulted from differences in opinion of when exactly the conversation had drifted too far from the topic to be relevant.

For the 14% of utterances with mismatched labels, we chose the label that would be “safest” in the information retrieval context where small talk might get discarded. If any of the annotators thought a given utterance was On-Topic, we kept it On-Topic. If there was a disagreement between Metaconversation and Small Talk, we used Metaconversation. Thus, a Small Talk label was only placed if all annotators agreed on it.

5 Experimental Setup

5.1 Features

As indicated in Section 1, we apply machine learning algorithms to utterances extracted from telephone conversations in order to learn classifiers for Small Talk, Metaconversation, and On-Topic. We represent utterances as feature vectors, basing our selection of features on both linguistic insights and earlier text classification work. As described in Section 2, work on the linguistics of conversational

Small Talk	Metaconv.	On-Topic
hi	topic	,
.	i	–
's	it	you
yeah	this	that
?	dollars	the
hello	so	and
oh	is	know
'm	what	a
in	was	wouldn
my	about	to
but	talk	like
name	for	his
how	me	they
we	okay	of
texas	do	't
there	phone	he
well	ah	uh
from	times	um
are	really	put
here	one	just

Table 2: The top 20 tokens for distinguishing each category, as ranked by the feature quality measure (Lewis and Gale, 1994).

speech (Cheepen, 1988; Laver, 1981) implies that the following features might be indicative of small talk: (1) position in the conversation, (2) the use of present-tense verbs, and (3) a lack of common helper words such as “it”, “there”, and forms of “to be”.

To model the effect of proximity to the beginning of the conversation, we attach to each utterance a feature that describes its approximate position in the conversation. We do not include a feature for proximity to the end of the conversation because our transcriptions include only the first ten minutes of each recorded conversation.

In order to include features describing verb tense, we use Brill’s part-of-speech tagger (Brill, 1992). Each part of speech (POS) is taken to be a feature, whose value is a count of the number of occurrences in the given utterance.

To account for the words, we use a bag of words model with counts for each word. We normalize words from the human transcripts by converting everything to lower case and tokenizing contractions

Features	Values
n word tokens	for each word, # occurrences
standard POS tags as in Penn Treebank	for each tag, # occurrences
line number in conversation	0-4, 5-9, 10-19, 20-49, >49
utterance type	statement, question, fragment
utterance length (number of words)	1, 2, ..., 20, >20
number of laughs	laugh count
n word tokens in previous 5 utterances	for each word, total # occurrences in 5 previous
tags from POS tagger, previous 5	for each tag, total # occurrences in 5 previous
number of words, previous 5	total from 5 previous
number of laughs, previous 5	total from 5 previous
n word tokens, subsequent 5 utterances	for each word, total # occ in 5 subsequent
tags from POS tagger, subsequent 5	for each tag, total # occurrences in 5 subsequent
number of words, subsequent 5	total from 5 subsequent
number of laughs, subsequent 5	total from 5 subsequent

Table 3: Summary of features that describe each utterance.

and punctuation. We rank the utility of words according to the feature quality measure presented in (Lewis and Gale, 1994) because it was devised for the task of classifying similarly short fragments of text (news headlines), rather than long documents. We then consider the top n tokens as features, varying the number in different experiments. Table 2 shows the most useful tokens for distinguishing between the three categories according to this metric.

Additionally, we include as features the utterance type (statement, question, or fragment), number of words in the utterance, and number of laughs in the utterance.

Because utterances are long enough to classify individually but too short to classify reliably, we not only consider features of the current utterance, but also those of previous and subsequent utterances. More specifically, summed features are calculated for the five preceding utterances and for the five subsequent utterances. The number five was chosen empirically.

It is important to note that there is some overlap in features. For instance, the token “?” can be extracted as one of the n word tokens by Lewis and Gale’s feature quality measure; it is also tagged by the POS tagger; and it is indicative of the utterance type, which is encoded as a separate feature as well. However, redundant features do not make up a sig-

nificant percentage of the overall feature set.

Finally, we note that the conversation topic is *not* taken to be a feature, as we cannot assume that conversations in general will have such labels. The complete list of features, along with their possible values, is summarized in Table 3.

5.2 Experiments

We applied several classifier learning algorithms to our data: Naive Bayes, Support Vector Machines (SVMs), 1-nearest neighbor, and the C4.5 decision tree learning algorithm. We used the implementations in the Weka package of machine learning algorithms (Witten and Frank, 2005), running the algorithms with default settings. In each case, we performed 4-fold cross-validation, training on sets consisting of three of the conversations in each topic (15 conversations total) and testing on sets of the remaining 1 from each topic (5 total). Average training set size was approximately 3800 utterances, of which about 700 were Small Talk and 350 Metaconversation. The average test set size was 1270.

6 Results

6.1 Performance of a Learned Classifier

We evaluated the results of our experiments according to three criteria: accuracy, error cost, and plausibility of the annotations produced. In all

Algorithm	% Accuracy	Cohen’s Kappa
SVM	76.6	0.44
C4.5	68.8	0.26
k-NN	64.1	0.20
Naive Bayes	58.9	0.27

Table 4: Classification accuracy and Cohen’s Kappa statistic for each of the machine learning algorithms we tried, using all features at the 100-words level.

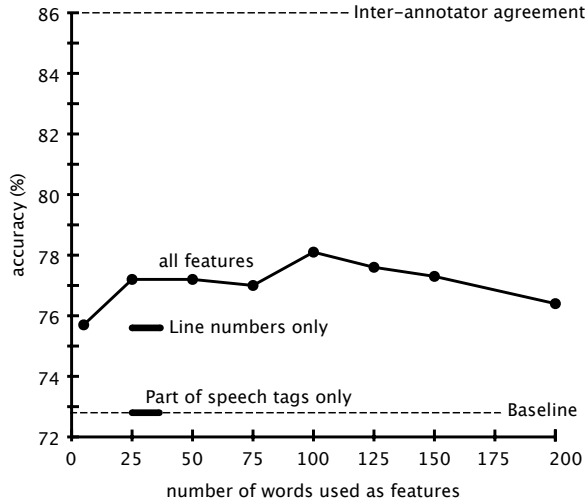


Figure 1: Classification results using SVMs with varying numbers of words.

cases our best results were obtained with the SVM. When evaluated on accuracy, the SVM models were the only ones that exceeded a baseline accuracy of 72.8%, which is the average percentage of On-Topic utterances in our data set. Table 4 displays the numerical results using each of the machine learning algorithms.

Figure 1 shows the average accuracy obtained with an SVM classifier using all features described in Section 5.1 except part-of-speech features (for reasons discussed below), and varying the number of words considered. While the best results were obtained at the 100-words level, all classifiers demonstrated significant improvement in accuracy over the baseline. The average standard deviation over the 4 cross-validation runs of the results shown is 6 percentage points.

From a practical perspective, accuracy alone is

S	M	T	<- classified as
55%	7%	38%	Small Talk
21%	37%	42%	Metaconv.
8%	3%	89%	On Topic

Table 5: Confusion matrix for 100-word SVM classifier.

not an appropriate metric for evaluating our results. If the goal is to eliminate Small Talk regions from conversations, mislabeling On-Topic regions as Small Talk potentially results in the elimination of useful material. Table 5 shows a confusion matrix for an SVM classifier trained on a data set at the 100-word level. We can see that the classifier is conservative, identifying 55% of the Small Talk, but incorrectly labeling On-Topic utterances as Small Talk only 8% of the time.

Finally, we analyzed (by hand) the test data annotated by the classifiers. We found that, in general, the SVM classifiers annotated the conversations in a manner similar to the human annotators, transitioning from one label to another relatively infrequently as illustrated in Table 1. This is in contrast to the 1-nearest neighbor classifiers, which tended to annotate in a far more “jumpy” style.

6.2 Relative Utility of Features

Several of the features we used to describe our training and test examples were selected due to the claims of researchers such as Laver and Cheepen. We were interested in determining the relative contributions of these various linguistically-motivated features to our learned classifiers. Figure 1 and Table 6 report some of our findings. Using proximity to the beginning of the conversation (“line numbers”) as a sole feature, the SVM classifier achieved an accuracy of 75.6%. This clearly verifies the hypothesis that utterances near the beginning of the conversation have different properties than those that follow.

On the contrary, when we used only POS tags to train the SVM classifier, it achieved an accuracy that falls exactly at the baseline. Moreover, removing POS features from the SVM classifier *improved* results (Table 6). This may indicate that detecting off-topic categories will require focusing on the words rather than the grammar of utterances. On

Condition	Accuracy	Kappa
All features	76.6	0.44
No word features	75.0	0.19
No line numbers	76.9	0.44
No POS features	77.8	0.46
No utterance type, length, or # laughs	76.9	0.45
No previous/next info	76.3	0.21
Only word features	77.9	0.46
Only line numbers	75.6	0.16
Only POS features	72.8	0.00
Only utterance type, length, and # laughs	74.1	0.09

Table 6: Percent accuracy and Cohen’s Kappa statistic for the SVM at the 100-words level when features were left out or put in individually.

the other hand, part of speech information is implicit in the words (for example, an occurrence of “are” also indicates a present tense verb), so perhaps labeling POS tags does not add any new information. It is also possible that some other detection approach and/or richer syntactic information (such as parse trees) would be beneficial.

Finally, the words with the highest feature quality measure (Table 2) clearly refute most of the third linguistic prediction. Helper words like “it”, “there”, and “the” appear roughly evenly in each region type. Moreover, *all* of the verbs in the top 20 Small Talk list are forms of “to be” (some of them contracted as in “I’m”), while *no* “to be” words appear in the list for On-Topic. This is further evidence that differentiating off-topic speech depends deeply on the meaning of the words rather than on some more easily extracted feature.

7 Future Work

There are many ways in which we plan to expand upon this preliminary study. We are currently in the process of annotating more data and including additional conversation topics. Other future work includes:

- applying topic segmentation approaches to our data and comparing the results to those we have obtained so far;

- investigating alternate approaches for detecting Small Talk regions, such as smoothing with a Hidden Markov Model;
- using semi-supervised and active learning techniques to better utilize the large amount of unlabeled data;
- running the experiments with automatically generated (speech recognized) transcriptions, rather than the human-generated transcriptions that we have used to date. Our expectation is that such transcripts will contain more noise and thus pose new challenges;
- including prosodic information in the feature set.

Acknowledgements

The authors would like to thank Mary Harper, Brian Roark, Jeremy Kahn, Rebecca Bates, and Joe Cruz for providing invaluable advice and data. We would also like to thank the student volunteers at Williams who helped annotate the conversation transcripts, as well as the 2005 Johns Hopkins CLSP summer workshop, where this research idea was formulated.

References

- Rebecca Bates, Patrick Menning, Elizabeth Willingham, and Chad Kuyper. 2005. Meeting Acts: A Labeling System for Group Interaction in Meetings. August.
- Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning*.
- Timothy Bickmore and Justine Cassell. 2000. How about this weather?: Social Dialogue with Embodied Conversational Agents. AAI Fall Symposium on Socially Intelligent Agents.
- Eric Brill. 1992. A simple rule-based part of speech tagger. Proc. of the Third Conference on Applied NLP.
- Christine Cheepen. 1988. *The Predictability of Informal Conversation*. Pinter Publishers, London.
- Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multiparagraph Subtopics Passages. *Computational Linguistics*, 23(1):33–64.
- John Laver, 1981. *Conversational routine*, chapter Linguistic routines and politeness in greeting and parting, pages 289–304. Mouton, The Hague.

LDC. 2004. Fisher english training speech part 1, transcripts. <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T19>.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. Proc. of SIGIR.

Jeffrey C. Reynar. 1999. Statistical models for topic segmentation. Proceedings of the 37th Annual Meeting of the ACL.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, San Francisco.