Automatic Identification of Off-Topic Regions of Conversation

by

Robin S. Stewart

Andrea Danyluk, Advisor

A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Cognitive Science

WILLIAMS COLLEGE
Williamstown, Massachusetts

May 22, 2006

Acknowledgements

The biggest thanks goes to my advisor, Andrea Danyluk, without whom this thesis clearly would not have been possible. Also crucial to its success were the timely and insightful suggestions of Yang Liu, who was always ready to give feedback and encouragement despite her physical location across the country. My second advisor, Joe Cruz, greatly helped in keeping us on-topic for a cognitive science thesis – at least, whenever he managed to show up. Kris Kirby contributed several important suggestions over cognitive science lunches, and deserves huge thanks for mentoring me over the years in classes and independent studies. I would also like to acknowledge Mary Harper, Brian Roark, Jeremy Kahn, and Rebecca Bates for providing invaluable advice and data. And I certainly owe more than iTunes gift certificates to the student volunteers at Williams who helped annotate the conversations that I used for this thesis.

Of course, in the bigger picture, I wouldn't be where I am without everyone out there who I love so much: mom, dad, sista, grandma "B", grandpa "Abe", Camille, my fellow Bhangra dancers, and all you Williams crock pots.

Last but not least, I am deeply grateful to the 2005 Johns Hopkins CLSP summer workshop, where I was introduced to the field of natural language processing and formulated this research idea.

Abstract

A collection of recorded and transcribed telephone conversations clearly demonstrates the universality of small talk and other socially-motivated utterances. Building on theories about the linguistics of conversational speech, I consider various ways of describing each utterance, including which words were used, their part-of-speech, and the proximity to the beginning of the conversation. In order to better understand which of these features are most useful, I create a system for automatically distinguishing between on- and off-topic utterances and compare its performance when using different combinations of these features. The central hypothesis is that conversational speech contains sufficient low-level clues to separate on- and off-topic utterances with an automatic classifier. I find that the overall structure of conversations is predictable, and automatic classification can indeed be done with better-than-chance accuracy. But distinguishing more reliably between on- and off-topic utterances will probably require deeper knowledge of the context and overall topic.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Spoken conversation contains noise on every level. The audio signal must be processed to separate vocal information from background noise and to extract real words. The sentences themselves are often fragmented, ungrammatical, and contain false starts, pauses, and corrections. Finally, the topics covered in the conversation may not connect from one utterance to the next, and may have nothing to do with the purported aim of the interaction. This thesis is concerned with the last category: identifying sections of a conversation that are irrelevant to the primary topic of conversation.

I have used a collection of recorded and transcribed telephone conversations [LDC04] as the basis for my investigations. These conversations clearly demonstrate the importance of small talk and other socially-motivated utterances. Although the participants of each conversation are assigned a topic to discuss, they frequently revert to other topics such as their hometowns, hobbies, significant others, and current events. They also sometimes talk about the conversation itself or the choice of topic. I refer to these phenomena as **small talk** and **metaconversation**, respectively.

The primary goals of this thesis are to better understand the distinctions between on- and off-topic utterances and to create a system for automatically detecting them. These two goals reinforce each other: better knowledge of the structure of each type of conversation will help to inform the process of crafting a good detection system; and the models created by automatic detection algorithms will in turn shed light on the characterization of on- and off-topic regions of conversation. To accomplish these goals, I combine theories about the linguistics of conversational speech with research in text and speech processing.

1.1 Motivation

Speakers appear to 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 [Che88]. In the context of the recorded telephone conversations, I approximate the distinction between transactional and interactional goals with the notion of rele-

vant and irrelevant topics of conversation. Because the participants of each telephone conversation have been given a specific topic to discuss for ten minutes, 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 all usually have specific topics of discussion.

In this thesis I hypothesize that utterances in conversational speech have sufficient low-level structure and verbal clues to separate out the different utterance types with an automatic classifier. One practical application of such a classifier is to more accurately search through sources of spoken data for information retrieval programs. For example, a search for information about weather patterns should not return conversations that include small talk like: "The weather's been cold recently, eh?" A search for information on sleeping patterns should not return a classroom lecture where the professor jokes about students falling asleep in class. In general, the ability to more deeply understand spoken language is useful to any system that seeks to analyze information with human-like skill.

1.2 Goals of This Thesis

The primary aim of this thesis is to determine how utterances differ depending on whether or not they are on-topic. More specifically, I distinguish between three types of utterance: on-topic, metaconversation, and small talk. As discussed more fully in Chapter 3, conversation is **On-Topic** if the conversants are discussing something at least tangentially related to the assigned topic; **Metaconversation** includes conversation about the assignment of the topic, the task itself, or administrative or technical details relating to the call; and **Small Talk** is defined to include everything else, i.e., conversation that is not even remotely related to the assigned topic.

Related work in text processing and linguistics suggests a number of features that might be useful for distinguishing among these types of utterances. The most apparent features are the words themselves; for example, the word "hi" typically occurs only in small talk regions in this corpus. Because some linguistic theories predict differences in verb tenses, pronouns used, and other grammatical indicators, I consider part-of-speech data from the conversations as well. And since introductions are known to be a predictable element of human conversations, I also include a feature measuring the proximity of each utterance to the beginning of the conversation. One of the major goals of this thesis is to determine which of these features actually help to characterize the utterance types I have defined.

A supporting goal is determining how to best combine these features to build a working detection system. This task is related to research in automatic document classification and topic segmentation, where the semantic content of entire documents is analyzed. It is also closely related to the task of **dialog act** (DA) detection, where every utterance is interpreted as an action with its own specific goals. To detect DAs,

a string of words must be segmented and classified according to each phrase's purpose in the conversation. In this thesis, I similarly treat different regions of conversations as containing fundamentally different types of information.

1.3 Approach

My approach for modelling Small Talk, Metaconversation, and On-Topic regions is to use machine learning techniques as an empirical framework for comparing various characterizations of these regions. Specifically, I analyze conversations at the level of individual **utterances**, which I define as segments of speech that are delineated by periods and/or speaker changes. For each utterance, I generate a variety of features that describe it in different ways. Because utterances are long enough to classify as a unit but too short to do so reliably without taking into account their context, I explore two different ways of making use of contextual information. First, I train a classifier to choose a label for each utterance, based on features of the current utterance as well as previous and subsequent utterances. I also explore the approach of initially segmenting the conversation into topically-coherent segments and subsequently classifying these segments.

1.4 Organization of This Thesis

This chapter has introduced the motivations for studying on- and off-topic segments of conversation as well as the goals and approach that will be taken in this thesis. In the next chapter, I provide an overview of previous research in related areas and note relevant hypotheses predicted by these sources. In Chapter 3 I describe the nature of the data, the definitions of each conversation type, and the annotation process. In Chapter 4 I describe the utterance features and machine learning algorithms used for the experiments in this thesis. In Chapter 5 I describe the specific experiments performed and present the results for each. Finally, in Chapter 6 I analyze the implications of these results for the characterization of on- and off-topic regions and conclude by outlining areas of future work.

Chapter 2

Related Work

This chapter explores previous research in areas related to off-topic detection. I first turn to theories of the structure of informal conversation to describe the linguistic and psychological motivations for this thesis. I also note the implications of these theories on feature selection. Next I describe semi-supervised and active learning techniques for making use of large unlabeled corpora of natural language text while minimizing the number of examples that need to be annotated.

In the last three sections I discuss approaches to text classification, topic segmentation, and methods for joint classification and segmentation. **Text classification** is the task of assigning a category to a given text document. Depending on the domain, this document may be anything from a news headline to a full-length dissertation. In supervised classifier learning, the categories are specified ahead of time and training examples for each are provided. The unsupervised version is called **clustering** because it requires grouping given documents into clusters that seem related in some way. By contrast, the **topic segmentation** task involves breaking up an article or longer document into segments with a cohesive topic or subject. Supervised segmentation uses training examples of similar segmented documents, while unsupervised segmentation algorithms aim to be general enough to segment any document.

The machine learning algorithms that I use in this thesis are Naive Bayes, the C4.5 decision-tree learning algorithm, support vector machines (SVMs), k-nearest neighbor (k-NN), and neural nets with backpropagation. These are mentioned throughout this chapter but are more fully described in Section 4.2.

2.1 The Structure of Informal Conversation

John Laver analyzed small talk and developed a theory about what motivates it and how it depends on the social situation, such as the relative social rank of the participants [Lav75]. He also looked at the phases that small talk typically goes through and verified the hypothesis that greeting and parting routines in particular are highly formulaic [Lav81]. In his theory, he divides small talk into sentences that are **self-oriented**, **other-oriented**, and **neutral**. The neutral category consists of

topics that strangers can discuss without presuming anything about social rank. In this category, "the syntactic structure of the phrases is typically abbreviated, which helps participants to recognize their phatic communion function." Characteristic phrases are: "Nice day," "What weather," "Frost tonight," "Nice party," "About time the trains were cleaned" [Lav81, p.301]. This suggests that common helper words such as "it", "there", and forms of "to be" may be missing from small talk regions, thus helping to identify them.

Christine Cheepen created a model for the higher-level structure of conversation [Che88]. She focuses on the interpersonal aspect of conversation rather than the semantics or syntax of utterances, and she claims that from this viewpoint, informal, spontaneous conversation underlies *all* conversation and is therefore worth studying in detail. She 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. Many conversations are predominantly motivated by one or the other of these objectives, but most contain some of each. For example, business transactions might begin and end with small talk that enforces social status even though the main body of conversation pertains to the work being done.

Encounters which have primarily interactional goals are organized in Cheepen's framework into a macro-structure of four parts. In the **introduction** phase, speakers greet each other, e.g. "hello", "how are you". During **speech-in-action** regions, the speech is related to the present physical world or the activity of chatting, e.g. "what lovely weather" or "it is so nice to see you". The majority of utterances are found in **story** regions, which are defined to include any recounting of events, feelings, or thoughts. Finally, the **closing** phase consists of formulaic endings to conversations, e.g. "it's been lovely, see you soon".

From Cheepen's interpersonal perspective, the task undertaken in this thesis can be partially viewed as detecting when participants switch between interactional and transactional modes of conversation. The transactional goal in the telephone conversations is clearly to discuss the assigned topic. Since this goal directly involves the conversation itself (in a similar way as the goal of a chat is to chat), metaconversation commonly arises as one of the forms of speech-in-action. There are also several interactional goals, such as establishing a social relationship (however brief) and maintaining a polite, non-offensive conversation. These goals are virtually always manifested in the introductions, but they also often drive speech-in-action regions. For example, asking about their partner's age or marital status helps to establish a social dynamic, and neutral comments about the weather or location help to fill awkward pauses. Presumably, closings also exhibit mostly interactional goals, but neither my nor Cheepen's data include good examples due to recordings being truncated.

Since the transactional goal is essentially the foundation of the conversations of this thesis, it makes up the bulk of the utterances and serves as inspiration for most of the story regions. One way to discuss a topic is to relate stories about one's experience with that topic, whether that experience is direct or indirect, physical or intellectual. Conversely, it does not seem appropriate to relate stories that are off-topic unless

it clearly satisfies an interactional goal. An example of such an exception is when the assigned topic is uncomfortable for strangers to talk about, so they purposefully avoid it. But overall, it seems reasonable to believe that story regions generally coincide with on-topic conversation. Further evidence for this mapping comes from the observation that speech-in-action regions tend to provide breaks between story regions in a similar way that small talk and metaconversation seem to appear in short bursts between on-topic regions when the conversants perhaps run out of things to say.

Thus, small talk seems to be the result of interactional goals and is found in introductions, closings, speech-in-action, and sometimes stories; on-topic sections follow transactional goals and may be manifested as stories; and metaconversation arises from the intersection of both types of goal and can be found as both speech-in-action and story. This motivates several interesting suggestions. First, since introductions seem to be a universal part of the structure of conversation, a system that automatically classifies utterances should recognize the high probability that each conversation begins with small talk. Second, since stories tend to involve the past or hypotheses about the future while speech-in-action deals with the present, the difference between these two categories, and thus the difference between on-topic and off-topic, might be manifested in the verb tenses in these sections. As discussed in Section 5.6, this latter hypothesis was found not to be the case.

Related work on **speech act theory** posits that with each utterance, a conversant is committing an action, such as questioning, critiquing, or stating a fact. There have been many attempts to automatically assign speech acts to various types of conversation, including phone conversations [WKNN97], internet chat rooms [TAJ04], and meetings [BMWK05]. The definitions of speech acts given by different authors depend on the context in which they are used. In [WKNN97], the corpus consisted of recorded brief telephone conversations in the domain of appointment scheduling. Warnke et al. attempted to automatically detect **dialog acts** which they grouped into 18 categories such as "accept", "suggest", and "request" and a total of 42 subcategories, each of which apply to a subset of the primary dialog act categories. These acts encompass segments of speech which are often just a few words long, so there is generally more than one act per sentence.

In some contexts, higher-level labels are defined. For example, Bates et al. [BMWK05] are developing a specification for labeling **meeting acts**, each of which spans one or more underlying dialog acts in the meetings domain. In this specification, rather than using categories and subcategories, meeting acts can simply be embedded within each other at most one level deep. For instance, there can be a "brainstorming" session within a "reporting" session, even though "brainstorming" can also be used as an independent top-level label. The label Bates et al. use which is most closely related to this thesis is "commentary". This includes both commentary about the meeting itself and commentary unrelated to the meeting, like my definitions of metaconversation and small talk, respectively. Unfortunately, in their initial study inter-annotator reliability was quite poor: even when small errors such as offsets by one dialog act were ignored, agreement was only 47%. On the other hand, about 20

different tags were being applied, and the decision of nesting labels versus keeping them at the top level is often ambiguous in practice. Bates et al. are in the process of determining which labels are most important and which are most reliable.

Another related approach comes from the study of a multimodal computer kiosk which has a monitor to visually display results but can also interpret human speech and generate speech in response [OSSB01]. **Off-talk** in this context is defined as utterances by the human that are not meant for the machine to interpret as a command. Oppermann et al. divide the phenomenon into two categories: read off-talk (ROT) and other (OOT). ROT refers to situations where the user is simply reading aloud instructions that appear on the computer display. Other types of off-talk include other forms of speaking to oneself, swearing, and talking to other people (although this was not allowed in the controlled study they analyzed). In an interesting sense, metaconversation is analogous to ROT in that the humans are verbally confirming the task at hand. In their study, Oppermann et al. found that about 10% of human speech was off-talk, and that the presence of quieter than usual speech was a good indicator for off-talk. They also found that the word "m-hm" occurred much more frequently in off-talk than in computer-directed speech.

A final strand of related research comes from the study of automatic conversational agents. In this work, the emphasis is on generating small talk rather than detecting it, in order to increase rapport between humans and a computerized real-estate agent [BC99, BC00]. The agent attempts to incorporate small talk via discourse planning that includes multiple goals for the conversation which are largely based on the work of Cheepen and Laver. The system apparently does not attempt to recognize and respond to user-driven small talk, which would clearly be a crucial component of a truly conversational agent. My research into detecting small talk could thus be used to improve such a conversational system.

2.2 Semi-Supervised and Active Machine Learning

Classifier learning algorithms use training data to create a classifier that can categorize new, unseen examples. In **supervised classifier learning**, categories are specified ahead of time and training examples for each are provided. In this thesis, I use supervised classifier learning to build a system that can automatically distinguish between on- and off-topic utterances. **Semi-supervised** machine learning techniques also aim to classify examples into specified categories, but they use both annotated and unannotated data for training. The advantage of this over pure supervised learning is the ability to make use of large sources of information without the expense of annotating every example. Assuming that none of the data have already been labeled, a subset of the many examples must be chosen for annotation. The simplest way to do this is to pick randomly. That random subset can be used to train a classifier which then classifies all of the unlabelled examples. All of the data can then be used to retrain the classifier, which then re-classifies the data, repeating the cycle until

the classifier converges on a set of answers. This was explored for text with a Naive Bayes classifier in [NMTM99].

Unfortunately, there is no guarantee that the randomly chosen annotated subset which the rest of the process is based on is actually representative of the rest of the data. To improve performance, it would be preferable to choose a set which includes as many representative types of examples as possible without being redundant. To accomplish this goal, **pool-based active learning** techniques iteratively choose examples from the pool of documents that seem most useful for improving the classifier. The process starts by training a classifier on a few randomly chosen annotated examples. The rest of the data are then analyzed in some way to choose a set of examples which seem to have the most potential for improving the classifier. Those examples are manually annotated, the classifier is retrained, and the process repeats.

Lewis and Gale [LG94] introduced the use of **uncertainty sampling** in choosing examples to be annotated. This sampling method only requires that the classifier being used can output the probability of category membership rather than just a hard decision. Such a classifier is used to classify all of the data in the pool and provide the corresponding probability estimates. The b examples for which this probability is the lowest (i.e., for which the classifier is most uncertain) are chosen for human annotation before the classifier is retrained and the process repeated. Lewis and Gale set up an experiment where news headlines were classified as being associated with a keyword or not. Each headline had only one keyword associated with it. Since very few headlines in the large collection actually match a given keyword, using a random sample of 1000 stories resulted in a disproportionate number of negative training examples for each keyword. Using an initial random sample of 3 stories and then using uncertainty sampling to choose 996 more produced better results than learning from all >300,000 possible training examples.

Alternatively, the **Query-by-Committee** (QBC) method finds uncertain examples by creating several plausible classifiers from the labeled examples and computing the disagreement between their classifications on some subset of unlabeled examples [MN98]. The disagreement measure is based on the distribution of classifications among the committee members. To avoid choosing outlier documents, which have high disagreement between classifiers but don't help classify other documents, the **density** of each document is also taken into account. This measure, which is calculated once at the beginning of the process, quantifies the similarity of each document to the rest of the corpus. The disagreement and the density are multiplied to produce a score for each document, and the top-scoring n documents are chosen for annotation, where n is the smallest batch size possible given the constraints of the labeling effort.

Roy and McCallum [RM01] took the QBC approach one step farther by actually estimating the learned classifier's expected error and then choosing documents which, when labeled, would most reduce the error. The majority of their paper is spent describing the optimizations needed to make such an algorithm tractable, including incremental retraining and using a random subset of documents for the expected error calculations. Their experimental results are quite good; comparable accuracy to other

algorithms is often achieved when only half as much data is annotated. The QBC approach has also been extended to support vector machines (SVMs) [TK00]. This is possible because SVMs can be incrementally retrained efficiently as new examples are added to the labeled training set.

2.3 Text Classification

Since a variety of corpora and subsets of corpora have been used to evaluate text classification approaches, it is difficult to pinpoint the relative effectiveness of algorithms by simply comparing published results. To test the whole range of corpora, Yang [Yan97] carried out new experiments with fairly simple categorization methods on common corpora, and ranked the classification algorithms accordingly. Yang concluded that corpus differences indeed significantly affect the algorithms' performance and noted that there is no evidence that more complicated techniques are in general better than simpler ones. She also found that the k-NN method is the most robust across corpora. Sebastiani [Seb02] provides a more recent comprehensive survey of text categorization definitions, approaches, evaluation metrics, and comparisons of published results. His strongest conclusion about the relative effectiveness of algorithms is that such comparisons are highly application-dependent.

Joachims [Joa98] showed that many n-gram features can significantly improve classifier performance, even though most of these n-grams provide little information individually and don't appear very often in test documents. SVMs can handle this large number of features well when most do not fire for a given test document. Joachims compared a SVM-based classifier with Naive Bayes, Rocchio, C4.5, and k-NN classifiers for categorizing news stories into keyword categories. He tried several types of SVM kernel functions — polynomials of various degrees and radial basis functions. Words that appear at least 3 times in the training set and that are not stop words (such as "and", "or", and "the") were used as features. The SVM classifier performed statistically significantly better than non-SVM approaches. Joachims noted that the SVMs took longer to train than all other classifiers tried except C4.5.

The most significant difference between the bulk of text classification work and my research is that most previous work classifies relatively large segments of text, usually on the order of the length of a news story. By contrast, some of my proposed approaches to off-topic detection require classifying single utterances, which provide much less information to a classifier. Including features from previous and subsequent utterances will be important for supplementing this sparse information. Despite the many text classification strategies presented in the literature, Roy and McCallum point out that: "Naive Bayes is not always the best performing classification algorithm for text, but it continues to be widely used for the purpose because it is efficient and simple to implement, and even against significantly more complex methods, it rarely trails far behind in accuracy" [RM01, p.3]. This is useful to remember, especially when the classification system is made more complex by combining it with other components such as segmentation.

2.4 Text Segmentation

Most algorithms for text segmentation employ some form of **lexical cohesion analysis** — a measure of semantic similarity between texts, which Stokes fully overviews [Sto04]. **Fine-grain** approaches, used in domains such as dialog act segmentation, are mostly based on selecting cue phrases (marker words like "well," "uh," "finally," "because," "also") via decision tree learners or similar methods, rather than n-gram language models. **Coarse-grain** approaches, which are used for subtopic segmentation in articles, are generally based on lexical cohesion, which includes word-repetition-based systems, such as TextTiling [Hea97] and C99 [Cho05]; and statistical word association systems based on word co-occurrence statistics.

The task of detecting on- and off-topic regions falls somewhere in between these two levels of granularity, because it focuses on utterances rather than phrases or documents. Perhaps combining the results of both approaches will be most useful; this suggests using a machine learning algorithm that can include many types of features. Such an approach is presented by Beeferman et al. [BBL99] for topic segmentation (though still on the full document level). They create many **cue word** features for each word, including whether that word occurs in the previous n or next n sentences, where a new feature is generated for each value of n in the range 1 to 10. They also generate **trigger pair** features, each of which consists of two words that tend to co-occur within segments. The presence of trigger pairs lowers the probability of a boundary, while the cue word features may increase or decrease the probability. Beeferman et al. used an exponential model to select the top 100 features for inclusion in the actual segmentation process. With these they achieved 12% and 19% error rates on broadcast news and Wall Street Journal stories, respectively.

A brief overview of pre-1999 topic segmentation techniques, some unsupervised, some supervised, and some based on dictionaries or thesauri is given in [Rey99]. One interesting section (3.4) discusses the structure of discourse near topic changes. Definite noun phrases or possessives followed by nouns are more likely to occur in sentences beginning a new topic, while the presence of pronouns provides some evidence that no topic change has recently taken place. The segmentation algorithm Reynar presents considers the 230 words (because that was the average topic segment length) preceding and following a putative topic boundary and classifies that point as a boundary or not. The model is based on the probabilities of words appearing zero times, once, and more than once in the regions before and after the proposed boundary. The system achieved precision and recall of about 60%. Impressively, it did almost as well on a Spanish corpus even though it was not retrained on Spanish data. Finally, better performance on an information retrieval task was achieved when using his system's boundaries than when using the human-annotated boundaries.

Utiyama and Isahara [UI01] built on Reynar's and others' work to create a method of unsupervised topic segmentation designed for summarization. The method is based on finding the maximum-probability segmentation via a minimum-cost graph-path algorithm. It statistically analyzes the distribution of words within a proposed segmentation, under the assumption that different topics have different word distributions

and are statistically independent of each other. When evaluated on an artificial corpus created by combining articles on different topics and testing whether the system could recreate the boundaries, their system performed statistically better than the best preceding system, C99 [Cho05].

Stokes also describes several evaluation metrics for text segmentation [Sto04]. These include: standard recall, precision, and f-measure; f_error, which allows a margin of error for the boundaries; P_k, which looks at every pair of units k units apart to see whether they are correctly separated (or not); and WindowDiff [PH02], which moves a fixed-length window across the text and counts missed or false boundaries within it.

2.5 Integrated Classification and Segmentation

As mentioned earlier, off-topic detection involves aspects of both segmentation and classification. Thus an important area of related research looks at how best to combine these tasks. The domain with properties most similar to off-topic detection is **dialog act** (DA) detection and variants thereof, which must both segment a word stream and then classify each segment as a dialog act type.

Ang et al. [ALS05] undertook an initial effort at automatic DA detection in the meetings domain using five DA types. The overall plan was to first segment and then classify, as a discrete two-step process. They used pause durations and a hidden-event language model for segmentation; for classification, they used a Maximum Entropy (MaxEnt) system with textual features only, including the DA's length, the first two and last two words, and the first word of the subsequent DA. They describe several evaluation metrics for segmentation, classification, and joint segmentation/classification. Segmentation results yielded an error rate of about 35% using reference words. Classification on reference segments and reference words gave about 20% error. When automatically segmented and then automatically classified, fully 75% of words were in an incorrect segment or classification. Giving the classifier information about surrounding units' classifications seemed to have little effect.

Zimmermann et al. [ZLSS05b] next attempted to jointly segment and classify the dialogue acts in meetings. Their first approach was to modify the segmentation system from (Ang et al. 2005) to also predict the DA type following any predicted boundary. Their second approach was modeled on Hidden Markov Model (HMM) part-of-speech taggers. Instead of a part-of-speech tag, the system assigns a DA type to each word, including a special type for the first word of a DA, which indicates the presence of a segmentation boundary. Zimmermann et al. also proposed a simple **DA error rate** metric which requires each unit to have correct boundaries as well as correct classification. Their results were slightly worse than the above results that used a sequential approach, but they stressed that all of these experiments were simply initial attempts that incorporate few of the possible features and machine learning techniques.

In a similar research effort, Warnke et al. [WKNN97] tried to detect dialog acts in a

corpus consisting of recorded brief telephone conversations in the domain of scheduling an appointment. Their initial method, like that of Ang et al. [ALS05], involved first segmenting the text and then classifying the resulting segments. Their second method approached the task as an A* search along the word string for the optimal segmentation and classification choices, based on the probabilities that had been assigned to boundaries and DA types by the segmentation and classification systems, respectively. This more integrated approach seemed to give negligible improvement over the previous sequential approach; rather, the quality of the language model was the most significant factor affecting results. Zimmermann et al. [ZLSS05a] also tried the A* search approach and similarly only achieved a minimal improvement over earlier approaches.

2.6 Further Reading

An overview of the linguistics of informal conversation can be found in [Che88]. Comprehensive background on machine learning techniques is given in [Mit97] and [WF05]. Finally, a thorough introduction to contemporary natural language processing techniques is [MS99].

Chapter 3

Data and Annotation

In this chapter I describe the data bank of telephone conversations from which conversations were drawn for this thesis, explain how subsets of the data were chosen, and discuss the method of annotating these conversations.

3.1 Data Selection

I started with human-transcribed telephone conversations from the Fisher data [LDC04]. In each conversation, participants who do not know each other are randomly connected and assigned a topic to discuss for ten minutes. If they complete the conversation, they are each sent a check for ten dollars (a fact which often comes up in the dialog).

In addition to the human transcriptions, additional transcriptions were previously generated for each conversation by using a speech recognition system. This automatically-generated text was then aligned with the human transcriptions using word alignment techniques. Often this two-step process works well, but there are some sections of conversation for which an alignment cannot be found. To include the possibility of using timing information garnered from automatic transcripts, I removed the 1180 conversations from the corpus for which more than 20% of utterances could not be aligned with the automatic transcripts. This left a set of 5727 conversations in 40 topics from which I proceeded to select data for my study.

There are a broad range of assigned topics in the corpus, spanning personal issues, current events, and philosophical questions. A list of the conversation topics used in this thesis appears in Table 3.1.

For my initial experiments I selected a set of 20 conversations: 4 randomly chosen from each of the 5 topics "computers in education", "pets", "terrorism", "censorship", and "bioterrorism". I will refer to this as **Set 1**. It contains 5070 utterances.

For later experiments, I used a set of 105 conversations: 7 randomly chosen from each of the 15 topics in Table 3.1, including all of the conversations in Set 1. This set, **Set 2**, contains 26,588 utterances.

Title	Topic
Computers in Education	What do each of you think about computers in education? Do they improve or harm education?
Pets	Do either of you have a pet? If so, how much time each day do you spend with your pet? How important is your pet to you?
Terrorism	Do you think most people would remain calm, or panic during a terrorist attack? How do you think each of you would react?
Censorship	Do either of you think public or private schools have the right to forbid students to read certain books?
Bioterrorism	What do you both think the US can do to prevent a bioterrorist attack?
Professional Sports on TV	Do either of you have a favorite TV sport? How many hours per week do you spend watching it and other sporting events on TV?
Affirmative Action	Do either of you think affirmative action in hiring and promotion within the business community is a good policy?
Computer Games	Do either of you play computer games? Do you play these games on the internet or on CD-ROM? What is your favorite game?
Foreign Relations	Do either of you consider any other countries to be a threat to US safety? If so, which countries and why?
Corporate Conduct in the US	What do each of you think the government can do to curb illegal business activity? Has the cascade of corporate scandals caused the mild recession and decline in the US stock market and economy? How have the scandals affected you?
Life Partners	What do each of you think is the most important thing to look for in a life partner?
Hobbies	What are your favorite hobbies? How much time do each of you spend pursuing your hobbies? Do you feel that every person needs at least one hobby?
Family	What does the word family mean to each of you?
Outdoor Activities	Do you like cold weather or warm weather activities the best? Do you like outside or inside activities better? Each of you should talk about your favorite activities.
Friends	Are either of you the type of person who has lots of friends and acquaintances or do you just have a few close friends? Each of you should talk about your best friend or friends.

Table 3.1: Topics used in this thesis. The first group of five is used in most of the experiments. The second group of five topics was chosen with the intuition that they would be easy to separate from small talk. The third group of five was chosen as potentially difficult to differentiate from small talk.

3.2 Definitions

The primary aim of this thesis is to determine how utterances differ depending on whether they are on- or off-topic. In order to do this, a clear definition of on- and off-topic regions is needed. As discussed in Section 2.1, 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 I assign them a special region type: **Metaconversation**. The purely irrelevant segments I call **Small Talk**, and the remaining segments are defined as **On-Topic**.

I split each conversation into **utterances** – segments of speech that are delineated by periods and/or speaker changes. An utterance can be as short as a single laugh or as long as an extended run-on sentence. This unit was chosen because clear shifts between region types usually do not occur within that short a span. So the task of distinguishing between on- and off-topic regions of conversation, as defined in this thesis, is to label each utterance in a conversation as belonging to one of the three categories On-Topic, Metaconversation, and Small Talk:

- I define conversation as **On-Topic** (sometimes abbreviated **T**) if the conversants are discussing something at least tangentially related to the assigned topic for the conversation. They need not be directly answering the questions posed in the topic description, so long as what they are talking about clearly follows from those questions.

- I define **Metaconversation** (**M**) as conversation about the assignment of the topic (e.g., "We're supposed to be talking about public education..."), conversation about the task (e.g., "How many of these calls have you done before?"), and conversation about administrative or technical details relating to the call (e.g., "I think we just wait until the robot operator comes back on the line.").

- I define **Small Talk** (**S**) to include everything else, i.e., conversation that is not even remotely related to the assigned topic. Some examples of this are: exchanging names ("I'm Michelle, nice to meet you."), locations ("So where are you calling from?"), living situation ("Oh, I live in a condo in Atlanta."), weather ("I hear it's pretty hot down there..."), and current activities ("I'm just sitting here rocking my baby.").

My definitions of these labels differ somewhat from their standard meanings. On-topic conversation does not have to actually be about the topic; it only has to be related to the topic. And even very intellectual conversation is labeled as small talk if it is not related to the assigned topic. Thus, small talk refers more broadly to utterances that are primarily motivated by interactional goals (as discussed in Section 2.1).

Label	Utterance
S	2: Well, hi there. [LAUGH]
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 3.2: An annotated conversation fragment.

3.3 Annotation

Annotation consists of manually assigning one of the three labels defined above (Meta-conversation, Small Talk, or On-Topic) to each utterance in a conversation. An excerpt from an annotated conversation appears in Table 3.2.

Each conversation in Set 1 was annotated by at least two people. Eight annotators, including myself, were involved in this effort. Annotators labeled conversations either on paper or with a web-based interface. The original conversation audio was not provided; all labeling decisions were made from the text transcriptions only. To make the process faster, only *changes* in label are marked; any lines left blank are assumed to retain the same annotation as the most recent previous label. The full annotation guide given to annotators appears in Appendix A. Note that I also explained the process and the definitions to all annotators in person. Moreover, for all of the conversations in Set 1, I was present in the room to answer any questions that came up during annotation. This sometimes included questions about the choice of label for a given utterance. I answered such questions by repeating the label definitions and urging the annotator to go with their best guess. For Set 2, I was only present for each annotator's first few conversations.

On average, pairs of annotators agreed with each other on 86.1% of the utterances in Set 1. The average of Cohen's Kappa statistic for each conversation was 0.70. 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. There were also occasional "outlier" conversations, such as one where a participant had a complete political agenda on top of the standard transactional and interactional goals. However, no conversations were post-hoc removed from the data sets.

The fairly high level of inter-annotator agreement overall was sufficient to continue annotating conversations for Set 2 without modifying the label definitions. Most

conversations had only one annotator, but as a quality check, eight of the new conversations in this larger set were randomly chosen to be annotated a second time. Annotators agreed with each other on 81.8% of the utterances in these eight conversations, with an average of Cohen's Kappa statistic of 0.67.

With multiple annotations of each conversation, I needed a way to deal with the 14% of utterances with mismatched labels. Because a practical result of this research might be the ability to use small talk detection for information retrieval of spoken "documents," the method I normally used was to select the label that would be "safest" under the assumption that small talk might get discarded. If any of the annotators thought a given utterance was On-Topic, I kept it On-Topic. If there was a disagreement between Metaconversation and Small Talk, I used Metaconversation. Thus, a Small Talk label was only placed if all annotators agreed on it.

Chapter 4

Experimental Setup

In order to determine the features that best characterize on- and off-topic regions, I apply machine learning algorithms to utterances extracted from telephone conversations in order to learn classifiers for Small Talk, Metaconversation, and On-Topic. To do this, I represent utterances as feature vectors, basing the selection of features on both linguistic insights and earlier text classification work. This chapter describes each feature in turn and then outlines the experimental setup I use to train and test classifiers.

4.1 Features

As described in Chapter 2, the work of [Lav81] and [Che88] on the linguistics of conversational speech 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, I number each utterance according to its line number: a label in the set {1-4, 5-9, 10-19, 20-49, "more than 49"}. I do not include a feature for proximity to the end of the conversation because the transcriptions include only the first ten minutes of each recorded conversation.

In order to include features describing verb tense, I use Brill's part-of-speech tagger [Bri92]. Each part of speech is taken to be a feature, whose value is a count of the number of occurrences in the given utterance. I used the standard part-of-speech tags from the Penn Treebank [San90], shown in Table 4.1.

To account for the words, I use a bag-of-words model with counts for each word. I normalize the words from the human transcripts by converting everything to lower

Tag	Description
$	dollar
“	opening quotation mark
”	closing quotation mark
(	opening parenthesis
)	closing parenthesis
,	comma
–	dash
.	sentence terminator
:	colon or ellipsis
CC	conjunction, coordinating
CD	numeral, cardinal
DT	determiner
EX	existential there
FW	foreign word
IN	preposition or conjunction, subordinating
JJ	adjective or numeral, ordinal
JJR	adjective, comparative
JJS	adjective, superlative
LS	list item marker
MD	modal auxiliary
NN	noun, common, singular or mass
NNP	noun, proper, singular
NNPS	noun, proper, plural
NNS	noun, common, plural
PDT	pre-determiner
POS	genitive marker
PRP	pronoun, personal
PRP$	pronoun, possessive
RB	adverb
RBR	adverb, comparative
RBS	adverb, superlative
RP	particle
SYM	symbol
TO	“to” as preposition or infinitive marker
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle or gerund
VBN	verb, past participle
VBP	verb, present tense, not 3rd person singular
VBZ	verb, present tense, 3rd person singular
WDT	WH-determiner
WP	WH-pronoun
WP$	WH-pronoun, possessive
WRB	WH-adverb

Table 4.1: The part-of-speech tags from the Penn Treebank [San90].

case and tokenizing contractions and punctuation. Some of the results of this tokenization can be seen in Table 4.2. I then rank the utility of words based on the feature quality measure presented in [LG94]. I chose this method because my goal is to classify utterances, and their method was devised for the task of classifying similarly short fragments of text (news headlines), rather than long documents.

To calculate the feature quality scores, I first find the **difference score** d for each word. The purpose of the difference score is to give higher weight to words which appear much more frequently inside a category than outside (or vice versa). For example, the word "topic" appears many times in Metaconversation regions, but rarely in On-Topic or Small-Talk regions. Thus, it has a high d_M score, where the subscript M stands for the comparison in frequency between Metaconversation (positive) regions and non-Metaconversation (negative) regions. These difference scores are calculated according to the following formula [LG94]:

$$d_i = \log \frac{\frac{c_{pi}+(N_p+0.5)/(N_p+N_n+1)}{N_p+w(N_p+0.5)/(N_p+N_n+1)}}{\frac{c_{ni}+(N_n+0.5)/(N_p+N_n+1)}{N_n+w(N_n+0.5)/(N_p+N_n+1)}}$$

where d_i stands for the difference score with category i as the positive region; N_p and N_n are the total number of tokens in the positive and negative regions, respectively; c_{pi} and c_{ni} are the counts of the given word in the positive and negative regions; and w is an estimate of the total number of word features desired. Because w did not have a significant effect on the difference scores and the actual number of word features varied often, for convenience I fixed w at 100 for all experiments, regardless of the actual number of words used. Since the objective is to distinguish among all three region types, I took the final difference score to be $d = \max d_i$, the maximum score found after each of the three categories was treated as the positive region for that word. The reasoning behind this choice is the intuition that words which strongly differentiate between *any* two region types should be useful for classification.

Next, I compute the frequency of each word in the full corpus of 5727 conversations (this corpus is described in Section 3.1). I used the whole corpus rather than the smaller, annotated, corpora in order to achieve a higher degree of statistical certainty; words that appear only rarely in the full corpus are probably the least reliable for characterizing on- and off-topic utterances.

Finally, I multiply the overall word frequency with the difference score d to get the **feature quality score** for each word. Thus, the highest-scoring words both are highly relevant for differentiating between categories and are likely to be useful because they appear often. I order the word list according to the feature quality score and use the top n tokens as features, where n varies in different experiments. Table 4.2 shows the most useful tokens according to this metric for distinguishing between the three categories in Set 1. Specifically, the words were assigned to column k if difference score d_k was the highest of the three d_i's. The 20 tokens with the highest feature quality in each column are displayed in the table. The ranked list of 50 tokens with the highest overall feature quality scores appears in Table 4.3 along with their d_i scores. These are precisely the word features used in experiments with $n = 50$.

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 4.2: The top 20 tokens for distinguishing each category in Set 1, as ranked by the feature quality measure [LG94].

I also include as features the utterance type (statement, question, or fragment), number of words in the utterance, and number of laughs in the utterance, as indicated in Table 4.4.

Because utterances are long enough to classify individually but too short to classify reliably, I also consider features of previous and subsequent utterances. More specifically, summed features are calculated for the p preceding utterances and for the s subsequent utterances. As described in Chapter 5, for most experiments $p = s = 5$.

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 the feature quality measure; it is also tagged by the part-of-speech tagger; and it is indicative of the utterance type, which is encoded as a separate feature. However, redundant features do not make up a significant percentage of the overall feature set.

Finally, note that the conversation topic is *not* taken to be a feature, as one cannot assume that conversations in general will have such labels. The complete list of features, along with their possible values, is summarized in Table 4.4.

Word	d_S	d_M	d_T
hi	252.968	0.022	0.006
.	1.473	1.158	0.702
,	0.999	0.913	1.039
–	0.691	0.944	1.321
topic	0.040	234.262	0.012
i	0.900	1.451	0.898
you	1.081	0.639	1.108
it	0.818	1.409	0.968
that	0.688	0.825	1.409
the	0.717	0.891	1.324
and	0.966	0.759	1.143
know	0.669	0.663	1.573
's	1.228	1.059	0.839
yeah	1.285	1.170	0.777
?	2.484	1.579	0.399
a	0.831	0.881	1.204
wouldn	0.026	0.029	42.031
hello	36.087	1.223	0.008
to	0.912	1.026	1.056
like	0.626	0.911	1.431
his	0.039	0.042	28.795
oh	2.629	1.832	0.353
they	0.668	0.895	1.385
of	0.764	0.817	1.318
't	0.688	0.926	1.338
'm	2.816	1.549	0.359
he	0.279	0.314	3.792
this	0.871	4.168	0.443
dollars	0.296	24.326	0.109
uh	0.583	0.379	2.101
so	1.137	1.303	0.809
um	0.642	0.959	1.374
in	1.173	0.430	1.147
is	1.100	1.690	0.725
put	0.050	0.054	22.177
just	0.796	0.387	1.631
what	0.933	1.873	0.758
was	0.379	1.559	1.342
think	0.319	1.147	1.779
about	0.872	2.753	0.610
my	1.762	1.139	0.612
but	1.102	0.733	1.045
talk	0.783	13.854	0.143
for	0.800	1.687	0.885
don	0.610	1.124	1.305
me	0.850	3.360	0.533
name	13.623	0.033	0.124
be	0.702	0.338	1.867
have	0.955	0.927	1.067
right	0.789	0.908	1.232

Table 4.3: The overall top 50 tokens in Set 1 as ranked by the feature quality measure [LG94]. Difference scores for each word are also listed, indicating, e.g., that "hi" appears much more frequently in Small Talk regions than elsewhere.

Features	Values
n word tokens	for each word, # occurrences
part-of-speech tags	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
number of laughs	laugh count
n word tokens in previous p utterances	for each word, total # occurrences
part-of-speech tags, previous p	for each tag, total # occurrences
number of words, previous p	total from p previous
number of laughs, previous p	total from p previous
n word tokens, subsequent s utterances	for each word, total # occurrences
part-of-speech tags, subsequent s	for each tag, total # occurrences
number of words, subsequent s	total from s subsequent
number of laughs, subsequent s	total from s subsequent

Table 4.4: Summary of features. Usually $s = p = 5$.

4.2 Machine Learning Techniques

In this thesis I use the Weka implementations [WF05] of five machine learning algorithms: Naive Bayes, the C4.5 decision-tree learning algorithm, support vector machines, k-nearest neighbor, and neural nets with backpropagation. A **Naive Bayes** classifier makes the simplifying assumption that all features are independent of each other, given class information. It can thus estimate the probability of an example being in a given category by multiplying together the individual probability estimates given by each of the example's features. The decision tree algorithm builds a decision tree that can make classification decisions by sequentially looking up attribute values at each node until a leaf specifying a category is reached. **Support vector machines** (SVMs) learn a linear separator between categories in an n-dimensional space by selecting **support vectors** — the training examples which are closest to the boundary — aiming to maximize the margin between the boundary and the support vectors. **K-nearest neighbor** (k-NN) techniques compare a given example to all of the training examples and assign a category based on the categories of the k most similar training examples. Last, I used **neural nets** based on multilayer perceptrons and trained with backpropagation.

I found that SVMs distinguished between the on- and off-topic regions much more accurately than the other machine learning algorithms (see section 5.3). Therefore, this is the technique I used for most of the experiments described in this thesis.

Chapter 5

Experiments and Results

In order to determine how to best characterize Small Talk, Metaconversation, and On-Topic, I use machine learning techniques as an empirical way of comparing the usefulness of each feature that might serve as an identifier. Classifiers are learned from labeled training sets and then run on separate test sets so that classification accuracy can be analyzed. In most cases, the labels to be learned are the human annotations of Small Talk, Metaconversation, and On-Topic. In some experiments, the objective is a simpler binary choice between Small Talk or not. In others, I approximated a segmentation approach to the task by labeling utterances as being a boundary between segments or not, where the definition of "boundary" varies between experiments. The resulting classifiers in this latter case look for features that indicate a change from one region to another, rather than indications of the regions themselves.

The experiments I performed are described below. The first three were primarily aimed at creating a better classifier, while the rest focused on discovering the factors that are important for distinguishing between on- and off-topic regions.

1. I compared five machine learning algorithms to show that SVMs are most effective in distinguishing between utterance types.

2. I varied the number of word tokens used as features input to a SVM in order to find out how many result in the best classifier.

3. To determine whether Metaconversation confounds the classifier, I tried classifying Small Talk versus Not Small Talk and found that it did not improve performance.

4. To determine which features best characterize on- and off-topic regions, I systematically varied feature sets and compared the accuracy of the resulting classifiers.

5. To find out whether any of the features could indicate *changes* in region type, I trained classifiers to label boundaries.

6. Finally, to better understand the effect of the conversation topics, I compared the performance of classifiers trained on different sized data sets.

5.1 Data Sets

As noted in Section 4.2, I used the implementations in the Weka package of machine learning algorithms [WF05], running the algorithms with default settings (except where noted)[1]. Since there are four conversations in each of the five topics represented in Set 1, for all experiments using that set I performed four-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 one from each topic (five total). The 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. Since 72.8% of the utterances in this set are On-Topic, a very simple classifier could correctly label 72.8% of utterances by labeling *all* utterances as On-Topic. I use this level of accuracy as a baseline with which to compare the performance of more sophisticated classifiers.

For experiments using Set 2, I performed 10-fold cross-validation over 105 conversations, 7 randomly chosen from each of the 15 topics in Table 3.1. I assigned conversations to folds such that all topics are represented in each fold and no more than two conversations on a given topic appear in the same fold. Here, the average training set size was 23,929 utterances and the average test set size was 2659 utterances. The distribution of labels was similar to the distribution in Set 1; here it was 23.4% Small Talk, 8.5% Metaconversation, and 68.1% On-Topic. Thus the baseline accuracy is 68.1%.

5.2 Evaluation Metrics

I evaluated the results of the experiments according to four criteria: accuracy, likelihood of being correct by chance, types of errors, and plausibility of the annotations produced. These measures were used first to determine which machine learning algorithm performed best, and then to help answer the question of which features are indicative of on- and off-topic regions.

Accuracy is simply the number of utterances classified correctly divided by the total number of utterances in the test set of conversations. To account for the statistical possibility of classifications being correct by chance, I use Cohen's Kappa statistic. To indicate the types of errors being made, I provide for many experiments the resulting confusion matrix (for example, see Figure 5.1). Finally, human annotations generally consist of large blocks of utterance types, rarely switching to a new type. To see whether automatically annotated conversations had this property, I took a qualitative look at some of the annotations produced.

[1] All experiments were run using the command line interface to Weka 3.4.

ML Algorithm	Accuracy	Kappa
support vector machine	**76.1 %**	**0.42**
neural net	71.1 %	0.35
1-nearest neighbor	67.5 %	0.30
decision tree learner	67.1 %	0.27
Naive Bayes	57.4 %	0.26
all-on-topic baseline	72.8 %	–

Table 5.1: Accuracy and Cohen's Kappa for five machine learning algorithms using an identical feature set.

5.3 Rationale for Using SVMs

As mentioned in Section 4.2, I evaluated five machine learning algorithms: Naive Bayes, the C4.5 decision tree learning algorithm, support vector machines (SVMs), k-nearest neighbor (k-NN), and neural nets with backpropagation. These were chosen because they cover a broad range of machine learning techniques. To find out which algorithm is most effective for automatically distinguishing between utterance types, I ran all algorithms with identical feature sets. These features were:

- 50 word tokens
- line number
- utterance type
- utterance length
- number of laughs
- aggregate word counts from the previous and subsequent 5 utterances

as described in Table 4.4. I used 50 word tokens because that was roughly the most that could be reliably handled by all of the machine learning algorithms without running out of memory. This resulted in 159 features generated for each utterance in Set 1. The results for each machine learning algorithm appear in Table 5.1. The support vector machine substantially outperformed all of the other techniques in both accuracy and Cohen's Kappa. Naive Bayes substantially underperformed all other techniques.

Because the results for 1-nearest neighbor were below the 72.8% baseline of labeling all utterances as On-Topic, and because the k value of 1 seemed overly limiting, I ran the same experiment using 2, 3, 5, and 7 as values of k. These results appear in Table 5.2. The reason that results are worse when $k = 2$ is probably that the two nearest neighbors often have different classifications, so the algorithm simply has to randomly choose between them. When k is odd, there is a much better chance of a

k	Accuracy	Kappa
1	67.5 %	0.30
2	60.4 %	0.26
3	70.7 %	0.34
5	71.6 %	0.35
7	73.2 %	0.37

Table 5.2: Accuracy and Cohen's Kappa for the k-nearest neighbor algorithm with different values of k.

majority classification, which is in turn more likely to be the correct label. The classifier with $k = 7$ performed above the baseline, but took much longer to compute than any of the other machine learning algorithms I tried, with the exception of neural nets.

5.3.1 Plausibility of automatic annotations

To determine how well the annotations produced by each machine learning algorithm resembled those assigned by human annotators, I manually looked over several conversations from each machine learning experiment and compared the human and machine annotations. The annotations produced by the SVM and 1-NN classifiers are shown along with the human annotations for a small excerpt of conversation in Table 5.3.

In my cursory analysis, I noted that 1-nearest neighbor annotations were very "jumpy" — the label rarely stayed constant for more than a few utterances in a row. Increasing the k value to 3 only moderately improved the plausibility of annotations. By contrast, the decision tree and support vector machine annotations were considerably more constant, rarely switching between labels. Instead, most of their errors were made in large blocks of utterances. For example, there were regions where the human annotators marked a section of Small Talk but the computer kept the whole block On-Topic. Also, it sometimes took a few utterances for the automatic classifier to "catch up" with human annotations; the transitions to new labels often came a few sentences after the human-annotated transitions. This is evident in the transition from Metaconversation (M) to On-Topic (T) in Table 5.3. Neural net and Naive Bayes output was not analyzed.

5.3.2 The advantages of SVMs

As we have seen, SVMs performed the same as or substantially better than all other machine learning techniques that I considered. The SVM was the only machine learning algorithm able to beat the baseline using this feature set, with the single exception of the extremely long-running 7-nearest neighbor classifier. In addition, it created plausible blocks of contiguously-labeled utterances at least as well as the

Human	SVM	k-NN	Utterance
S	S	S	2: Well, hi there. [LAUGH]
S	S	S	2: [LAUGH] Hi.
S	S	S	2: How nice to meet you.
S	S	S	1: It is nice to meet you too.
M	M	M	2: We have a wonderful topic.
M	M	M	1: Yeah.
M	M	T	1: It's not too bad. [LAUGH]
T	M	M	2: Oh, I — I am one hundred percent in favor of, uh, computers in the classroom.
T	M	M	2: I think they're a marvelous tool, educational tool.
T	T	M	2: Uh, how do you feel about it?
T	T	T	1: Yeah.
T	T	T	1: Actually I'm not really too familiar with it.

Table 5.3: A conversation fragment comparing human, SVM, and k-NN annotations.

S	M	T	<– classified as
55%	4%	41%	(S)mall Talk
15%	**27%**	58%	(M)etaconv.
7%	5%	**88%**	On-(T)opic

Figure 5.1: Confusion matrix for the SVM classifier of Section 5.3.

other algorithms I explored. Finally, the confusion matrix for the SVM using the feature set described above appears in Figure 5.1. This matrix shows that the SVM was able to detect a significant number of each utterance type: 55% of the Small Talk, 27% of Metaconversation, and 88% of On-Topic utterances. In addition, I note that it incorrectly labels On-Topic utterances as Small Talk only 7% of the time. This is good news for potential information retrieval applications which might automatically delete small talk sections and thus seek to minimize this type of error. For all of these reasons, I will only report the performance of SVMs in further experiments.

5.4 The Impact of Words

A major linguistic and practical question that needed answering was how many words the classifier should consider. From the perspective of learning how to characterize the different utterance types, I sought to find out whether only a few key words were relevant, or if ever increasing numbers of words contain useful information. From a more pragmatic perspective, I needed to find the number of words that would result in the best performance while remaining tractable — computing in a reasonable length

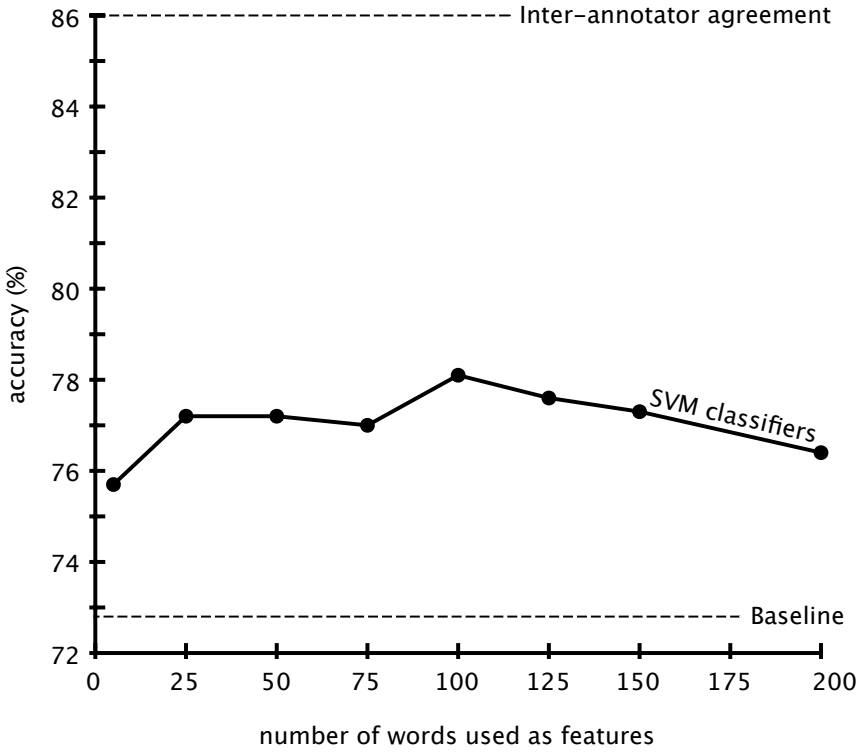

Figure 5.2: Classification results using SVMs with varying numbers of words.

of time and without running out of memory. To find out, I trained SVM classifiers on Set 1 with the features:

- n word tokens
- line number
- utterance type
- utterance length
- number of laughs
- aggregate word counts from the previous and subsequent 5 utterances

where n, the number of words, was the only parameter that varied between experiments.

The results appear in Figure 5.2. The classifier with 100 word features attained the highest accuracy (78.1%), while classifiers with fewer and more word features seemed to do progressively worse. It may be that only 100 words are useful for distinguishing between region types, but it seems more likely that this maximum is a result of data

S	N	<– classified as
48.2%	51.8%	**S**mall Talk
6.9%	93.1%	**N**ot Small Talk

Figure 5.3: Confusion matrix for a SVM classifier of Small Talk versus Not Small Talk.

sparseness — the classifier does not have a reliable estimate of which region type is predicted by less common words. So as to use the presumed best system possible for making comparisons, I used 100 word features in most subsequent experiments.

5.5 Distinguishing On-Topic from Metaconversation

Metaconversation is somewhat of an outlier class. It occurs in only 9% of utterances, and since much of it comprises conversation *about* the topic, it might be reasonable to imagine that it would contain many of the same words as On-Topic conversation. It may even be the case that distinguishing between Metaconversation and On-Topic is not useful for understanding interactional speech or for improving applications in information retrieval. Most importantly for this thesis, in order to determine with any confidence whether specific features are important or not and whether there is a clear structure to small talk, it is necessary to have as accurate a classifier as possible.

To find out whether including Metaconversation as a separate class was positively or negatively affecting classification accuracy, I performed binary classification experiments to automatically distinguish between Small Talk and Not Small Talk. In this case the baseline is formed by classifying all utterances as Not Small Talk, which yields an accuracy of 82.1%. I once again used the data in Set 1 and the same feature set as for the experiments comparing machine learning techniques (Section 5.3): 50 word tokens, the line number, utterance type, utterance length, number of laughs, and aggregate word counts from the previous and subsequent 5 utterances.

Under these conditions, the accuracy was 85.1% and Cohen's Kappa statistic was 0.45. The full confusion matrix is shown in Figure 5.3. It is clear from Table 5.4 that this is not substantially different from the results for the 3-way SVM classifier. The Kappa statistic and percentage gain in accuracy over the baseline are roughly the same for each setup. From this I conclude that Metaconversation is indeed an independent classification which does not detract from the ability to detect Small Talk.

Condition	Baseline	SVM Accuracy	% Gain	Kappa
S/M/T	72.8%	76.1%	4.5%	0.42
S vs. Not S	82.1%	85.1%	3.7%	0.45

Table 5.4: Comparison of the performance of a 3-way Small Talk / Metaconversation / On-Topic classifier and a binary Small Talk vs. Not Small Talk classifier.

5.6 Relative Utility of Features

The primary goal of this thesis is to determine how well various features characterize on- and off-topic regions of conversation. To do this, I analyzed the relative contributions of each feature in creating effective learned classifiers. Several of these features were selected due to the claims of linguists [Lav81, Che88]:

- position in the conversation (line numbers),
- verb tense (parts-of-speech),
- presence of common helper words such as "it", "there", and forms of "to be" (word features).

The other features were:

- utterance type,
- utterance length,
- number of laughs.

All of these are more fully described in Table 4.4.

Using the data in Set 1, I trained SVM classifiers with three different types of feature sets:

(a) all of the features listed above (to serve as a reference point);

(b) all of the features except one;

(c) one feature at a time.

I used 100 word features because it yielded the best accuracy overall, as reported in Section 5.4. The results of this systematic variation of feature sets appears in Table 5.5.

When proximity to the beginning of the conversation ("line numbers") is the sole feature, the SVM classifier achieves an accuracy of 75.6%. This clearly verifies the hypothesis that utterances near the beginning of the conversation have different properties from those that follow. The low Kappa statistic reflects the fact that *only* utterances near the beginning of the conversation were marked as Small Talk,

	Condition	Accuracy	Kappa
(a)	All features	76.6	0.45
(b)	No word features	75.0	0.20
	No line numbers	76.9	0.45
	No part-of-speech features	77.8	0.47
	No utterance type, length, or # laughs	76.9	0.46
	No previous/next info	76.3	0.22
(c)	Only word features	**77.9**	**0.47**
	Only line numbers	75.6	0.17
	Only part-of-speech features	72.8	0.00
	Only utterance type, length, and # laughs	74.1	0.09

Table 5.5: Percent accuracy and Cohen's Kappa statistic for the SVM at the 100-words level when features were (a) all used, (b) withheld one at a time, and (c) used individually.

S	M	T	<– classified as
17%	0%	83%	(S)mall Talk
4%	0%	96%	(M)etaconv.
0%	0%	100%	On-(T)opic

Figure 5.4: Confusion matrix for a SVM classifier using line number as the only feature.

and all of the Small Talk further down was mislabeled as On-Topic. In addition, no utterances were labeled as Metaconversation. The full confusion matrix appears in Figure 5.4.

By contrast, when I used only part-of-speech tags to train the SVM classifier, it achieved an accuracy that falls exactly at the baseline. A look at the confusion matrix confirms that indeed, all but 3 utterances are simply labeled as On-Topic, the majority class, indicating that part-of-speech features contain no useful information, at least for a SVM classifier. In fact, they seem to confuse the classifier. When I removed only part-of-speech tags from the otherwise full-featured SVM classifier, this actually *improved* results (Table 5.5). This may indicate that detecting off-topic categories will require focusing on the words rather than the grammar of utterances. Put another way, I have not found evidence that people subconsciously change the structure of their speech based on whether or not it is On-Topic. However, it is possible that some other detection approach and/or richer syntactic information (such as parse trees) would be beneficial to a classifier.

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 5.6: The top 20 tokens for distinguishing each category in Set 1, as ranked by the feature quality measure [LG94].

The words with the highest feature quality measure (Table 5.6) clearly refute the linguistic prediction that a lack of common helper words might be indicative of small talk. Instead, 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.

Finally, utterance type, utterance length, and number of laughs did not appear to help when combined with the other features, but unlike part-of-speech tags, they were able to label some Small Talk when used on their own. The confusion matrix for the experiment with utterance type, utterance length, and number of laughs as the only features appears in Figure 5.5. When each of these three features was used individually without the presence of the other two, no Small Talk or Metaconversation was labeled.

S	M	T	<– classified as
9.3%	0%	90.7%	(S)mall Talk
0.8%	0%	99.2%	(M)etaconv.
0.4%	0%	99.6%	On-(T)opic

Figure 5.5: Confusion matrix for a SVM classifier using utterance type, utterance length, and number of laughs as the only features.

5.7 Detecting Boundaries Between Regions

Detecting Small Talk, Metaconversation, and On-Topic utterances can also be viewed as a segmentation of conversation into relevant and irrelevant parts. Although features such as part-of-speech tags were not able to distinguish between utterance types, it is possible that they and other features might be able to characterize the *changes* between on- and off-topic regions. To find out whether this is the case, I converted the human annotations into the labels **Boundary** and **No Boundary**. The classifier's job is thus to automatically detect these boundaries based on the utterance features.

In these experiments, utterances were described by the following features (as in Section 5.3):

- 50 or 100 word tokens
- line number
- utterance type
- utterance length
- number of laughs
- aggregate word counts from the previous and subsequent 5 utterances

I varied the boundary size and number of word tokens considered to see if it would make any difference. In the first set of experiments, a boundary is defined as any utterance which is labeled differently from the previous utterance. In order to better simulate gradual changes between topics, in the second set of experiments a boundary includes any utterance that is within two utterances of a label change. For isolated label changes, this creates boundary regions with a width of five utterances. In both of these conditions I set the number of word tokens at both 50 and 100.

The results for training SVM classifiers on the conversations in Set 1 are shown in Table 5.7 along with the baseline accuracies, which are based on the accuracy of classifying all utterances as non-boundary. In each case, more boundaries are found than would be predicted by chance, though still not a very large percentage. Since many topic segmentation systems use part-of-speech features as clues of topic shifts, I also tried adding the part of speech tags back to the feature set. This did not seem

B. width	Words	Accuracy	B.'s found	Kappa
1	baseline	96.9 %	–	–
1	50	97.4 %	23.9 %	0.36
1	100	96.8 %	23.1 %	0.30
1	50+POS	97.4 %	16.2 %	0.26
5	baseline	87.8 %	–	–
5	50	89.4 %	22.9 %	0.31
5	100	87.9 %	30.3 %	0.32
5	50+POS	89.0 %	24.8 %	0.32

Table 5.7: Accuracy, percentage of boundaries (B.'s) found, and Cohen's Kappa for detecting boundaries between segments, under various conditions.

to significantly change the results (Table 5.7). When only part of speech tags were used as features, the classifiers did not find *any* boundaries. It is possible that the biggest problem here was the small number of boundaries relative to non-boundaries, leading to a classifier without enough confidence to label utterances as boundaries. This could be investigated in the future by balancing the number of training examples in each class.

5.8 Experiments with Larger Data Sets

A final important question is: to what extent do these characterizations of utterance type depend on the topics of conversation? Do they hold up under a greater variety of topics and a larger data set size? To find out, I compared the results of Set 1, containing 20 conversations on five topics, with Set 2, which has 105 conversations on 15 topics. Both sets are detailed in Section 5.1. Due to the large size of Set 2, 20 word tokens was the most that could fit in memory. Otherwise, I used the same basic feature set used in many of the experiments above: the line number, utterance type, utterance length, number of laughs, and aggregate word counts from the previous and subsequent 5 utterances. The summarized SVM classifier results are shown in Table 5.8, and the confusion matrices are in Figure 5.6.

Not surprisingly, the confusion matrix for Set 1 with 20 word tokens (Figure 5.6a) is very similar to that with 50 word tokens (Figure 5.1). The main difference that appears when moving to the larger variety of topics in Set 2 (Figure 5.6b) is the dramatically lowered ability to detect Small Talk and Metaconversation. This is probably a result of the larger number of topics making it more difficult for the classifier to predict whether or not an utterance is off-topic. For example, conversation about significant others is always labeled as Small Talk in Set 1, but since Set 2 includes a topic on life partners (see Table 3.1), it is very difficult for the classifier to predict whether utterances with words such as "spouse" are Small Talk or On-

Set	Baseline	Accuracy	Kappa
Set 1	72.8 %	77.2 %	0.40
Set 2	68.1 %	71.6 %	0.20

Table 5.8: Accuracy and Cohen's Kappa for classifiers trained on Sets 1 and 2, using identical features.

(a) Set 1:

S	M	T	<– classified as
46%	4%	50%	(S)mall Talk
15%	20%	65%	(M)etaconv.
6%	2%	92%	On-(T)opic

(b) Set 2:

S	M	T	<– classified as
15%	1%	84%	(S)mall Talk
9%	13%	77%	(M)etaconv.
1%	1%	98%	On-(T)opic

Figure 5.6: Confusion matrices for classifiers trained on (a) Set 1 and (b) Set 2, using identical features.

Topic because the classifier is never given the assigned topic. This may indicate that knowledge of the assigned topic is necessary for accurate detection of on- and off-topic regions. It is also likely that a classifier with the ability to analyze more than 20 word features would have better performance on Set 2.

5.9 Summary of Results

I find that conversations are structured in a way that some small talk can be identified without deep analysis. Words are the most important indication of utterance type, but it is not clear how to determine theoretically which words should indicate which region types. Information about the content of previous and next utterances is also important. But SVM classifiers do not seem to benefit from access to shallow features such as part-of-speech tags, which may be partly due to the fact that most of these other features are derived directly from the words, so are implicit in them. Finally, increasing the number of topics reduces the ability of SVM classifiers to distinguish between utterance types. This seems to indicate that knowledge of the assigned topic may be necessary for more reliable classification.

Chapter 6

Analysis and Future Work

This thesis has been concerned with determining how conversational utterances differ depending on whether they are on- or off-topic. The central hypothesis was that utterances in conversational speech have sufficient low-level structure and verbal clues that different utterance types might be identified with an automatic classifier. In particular, I was interested in the information that can be gleaned from the presence or absence of specific words, parts-of-speech, and the position of an utterance within the larger conversation.

I started by defining Small Talk, Metaconversation, and On-Topic classifications of utterances within a conversation. Annotators were able to label conversations according to this specification with a high degree of agreement. I then generated a variety of features for each utterance in two sets of conversations and used support vector machines to automatically learn classifiers. By analyzing the performance of classifiers with access to differing feature sets and data sets, I aimed to uncover what makes an utterance Small Talk, Metaconversation, or On-Topic.

6.1 Implications for Linguistic Hypotheses

Research on the linguistics of conversational speech led me to hypothesize 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”.

My findings, which supported only the first hypothesis, were as follows:

1. As expected, conversations in my data set have a predictable structure in that they routinely start with small talk, often followed by conversation about the topic. A classifier with no other information except line number labeled 17% of the small talk in the ten-minute conversations. This small talk serves the important goal of setting up an interpersonal context for the conversation.

2. Contrary to my hypothesis, part-of-speech tags do not appear to contain useful information for distinguishing between Small Talk, Metaconversation, and On-Topic regions. Classifiers using part-of-speech tags as the only features did not find a meaningful percentage of small talk, nor were classifiers improved when part-of-speech tags were added to other feature sets. This may indicate that detecting off-topic categories will require focusing on the words rather than the grammar of utterances; put another way, I have not found evidence that people subconsciously change the structure of their speech based on whether or not it is on-topic. However, the words themselves do carry some part-of-speech information, so it may not be so surprising that part-of-speech tags did not add new knowledge. Also, it is possible that some other detection approach and/or richer syntactic information (such as parse trees) would be beneficial to a classifier.

3. The types of words that were useful for distinguishing amongst categories (Table 4.2) did not uphold the hypothesis that a lack of common helper words might be indicative of small talk. Still, many of the words that were automatically chosen make intuitive sense as being important; for Small Talk, "hi", "hello", and "name" seem particularly obvious, while Metaconversation clearly favors "topic", "talk", and "phone". The choices of On-Topic words are less intuitive. For instance, from Table 4.3 we see that the tokens "wouldn", "his", and "put" appear almost exclusively in On-Topic utterances. These words do not present a clear pattern, and may represent the effects of data sparseness.

6.2 Other Findings

Utterance type, utterance length, and number of laughs are probably not very important features, but they do capture some useful information. Classifiers were useless when these features were considered individually, but when used together they were able to identify 9% of Small Talk.

Including information about the context of an utterance is clearly important. The Kappa statistic is twice as high when summary features from previous and next utterances are included than when they are not (Table 5.5). This is probably because a single utterance in isolation, sometimes including only one or even no words, contains far too little information to reliably classify. For example, a sentence such as "[LAUGH] I know." could appear in any of the region types. The best way to find out which one is correct is to look at the surrounding context. The only time that context may be a confounding input is near region boundaries; but boundaries are relatively infrequent.

Metaconversation is a valid utterance type in the telephone conversations I analyzed, distinct from the other utterance types. This was confirmed both by the ability of classifiers to detect some of it, and the lack of improvement in classifier performance when they were reduced to only distinguish between Small Talk and Not Small Talk.

Including more conversation topics makes it more difficult to distinguish between

utterance types, as seen in Section 5.8. This is true even though the set that included more topics also included more conversations in each individual topic. This suggests that knowledge of the topic may be necessary for fully characterizing the different types of utterances in a given conversation. Indeed, it would be very surprising if such knowledge is *not* necessary, given how integral it is to the human annotation specification (Appendix A).

Finally, it is notable that the classifier that only considered words as features outperformed all other classifiers. This confirms the idea, present too in the discussion above, that the words themselves remain the most important factor. Although I have been able to extract certain features that seem to be informative, such as the position in the conversation, I have not found any "magic" features that can approximate the membership in an on- or off-topic region. In the absence of such a feature, the words themselves remain crucial for automatic detection systems. It seems likely that classifiers which can analyze more word features and train on larger data sets will improve over the classifiers presented here, but this does not provide any deeper insight into how to represent the "meaning" of off-topic. Figuring out how to represent meaning is still very much an open problem.

6.3 Future Work

There are many ways to expand upon the research in this thesis. Annotating more data and including additional conversation topics would probably improve classification results as well as improve the statistical significance of those results. As discussed in Section 2.2, semi-supervised approaches for making use of the full unlabeled data set might be a good way to go about this. More varied data would also give a richer context for further investigation.

There are several candidate features that could potentially characterize on- and off-topic regions better than those considered in this thesis. They include:

- Parse structure — the full syntactic structure of each utterance. This would provide deeper information than that given by part-of-speech tags.
- Timing and pause duration. The length of pauses and the speed of talking may provide important clues about whether a speaker is off-topic.
- Prosodic information — dynamic descriptions of voice quality, such as tone and pitch. For instance, intuitively, speakers might sound more "conversational" in some measurable way when they know they are off-topic.

There are also several possibilities for increasing the accuracy or generality of classifiers.

- Because sentences are long enough to classify individually but too short to do so reliably without taking into account previous and subsequent sentences, the

automatic labeling task can be approached in several alternate ways. A similar approach to the one I followed is to determine for each sentence the probability of the most likely label, and then label the sentences where this probability is sufficiently high; the rest of the sentences can be classified using the nearest preceding high-probability label. Another possible approach is to classify each sentence based on a smaller number of features and then use a dynamic Bayesian network to smooth the classification.

- In general, spoken data does not come with human transcriptions. Thus, it would be interesting to run the experiments with automatically generated (speech recognized) transcriptions, rather than the human-generated transcriptions used in this thesis. Since automatic transcriptions contain more noise than those provided by humans, this will presumably pose new challenges.

Most importantly, the conversational telephone speech analyzed in this thesis is a fairly contrived and unusual form of spoken data, serving mainly as a useful test bed for initial study of the nature of small talk, the concept of interactional speech, and even the very meaning of relevance. Broadening the scope of this analysis to additional genres — such as broadcast news, broadcast conversations, meetings, and class lectures — is the ultimate goal of this research.

Appendix A

Annotation Guide

Off-Topic Detection: Annotation Guide

Robin Stewart (06rss_2@williams.edu)

November 19, 2005

I. Introduction

The data we are annotating are telephone conversations that have been recorded and transcribed. In each case, a volunteer calls an automated system which then calls the phone number of another volunteer in the database. If that person answers the phone, the system assigns them a topic to talk about with each other for ten minutes, after which they are mailed a check for ten dollars. The transcriptions we are using begin right after the computer system finishes telling the volunteers what the topic is. The speakers (total strangers to each other) do not usually spend all their time talking about the actual topic. They often spend time talking about the phone call itself, or talking *about* talking about the topic – a phenomenon we will call "metaconversation". Conversants also engage in small talk that is not related to the assigned topic, such as asking about the weather or imparting personal details.

The goal of the present research is to build a system that can automatically detect these regions of conversation. In order to "teach" such a system to recognize these regions, some conversations need to be tagged by human annotators. The conversations have been segmented into sentence-length or shorter utterances, each of which can be classified as either metaconversation (M), small talk (S), or "at least vaguely on-topic" (T). The conversations were transcribed quickly and there may be spelling, punctuation, or other errors, all of which you can ignore. Laughs, unidentifiable noises, and inaudible speech are indicated in [brackets].

II. Tag Descriptions

S: Small Talk

Conversation that is not even remotely related to the assigned topic. This includes:

- exchanging names ("I'm Michelle, nice to meet you.")
- exchanging locations ("So where are you calling from?")
- living situation ("Oh, I live in a condo in Atlanta.")
- weather ("I hear it's pretty hot down there...")
- sports ("Yeah, but it didn't stop the Braves from finishing out the ninth inning...")
- family or friends ("My girlfriend's from there so I've visited a few times.")
- current activities ("I'm just sitting here rocking my baby.")

M: Metaconversation

- Conversation about the assignment of the topic (e.g. "We're supposed to be talking about public education...")
- Conversation about the task: getting paid to talk on the phone to some stranger about some topic (e.g. "How many of these calls have you done before?")
- Conversation about administrative or technical details relating to the call (e.g. "I think we just wait until the robot operator comes back on the line.")

T: On-Topic

- Conversation that is at least vaguely or tangentially related to the assigned topic. These are regions that could plausibly be representative of "two people talking about topic X". Conversation is on-topic even if the conversants are not answering the questions posed in the topic description, so long as they are talking about something remotely related to or following from those questions.

Under this definition, **most conversations are mostly on-topic**. (And everything else is metaconversation or small talk.)

Some "sentences" consist of only a laugh or an "uh-huh", whose conversation type is unclear. These sentences should always be classified as the *preceding* region's conversation type. In other words, only mark transitions in conversation type next to the first sentence that makes it clear that the transition has taken place.

III. Labeling Format and Instructions

Before each conversation you will see the topic that was assigned and a small amount of information about speaker 1 and speaker 2. Each sentence or sentence fragment of the conversation appears in order on its own line, prefixed by the line number and then the speaker number (1 or 2). Finally, at the end of each conversation, there is space to include five keywords which describe what the on-topic regions were about.

Your job: Place a tag (**M**, **S**, or **T**) every time a change in conversation type takes place. If you are using a printout, jot the tag next to the line number. If you are annotating online, click the appropriate button next to the line. All unmarked lines are assumed to retain the same conversation type as the closest tag above.

When unsure: You must choose a conversation type to assign, but if you are unsure of your decision, **place a 'U' after the main tag**. Thus if you are unsure but your best guess is metaconversation, your label should be **MU**.

After reading each conversation: Briefly look back at the **on-topic regions only** and specify exactly five keywords to describe the subtopics addressed. Think of these keywords as search terms you would use if you wanted to find more information about the subtopics discussed in the on-topic regions.

V. Example Excerpt of an Annotated Conversation

Sample annotations appear on the left, and possible keywords for this excerpt appear at the end. Everything else is provided by the system. Note that the conversations you will be annotating are much longer than this excerpt.

Speaker 1: American female
Speaker 2: American male
Topic: According to each of you, which is worse: gossiping, smoking, drinking alcohol or caffeine excessively, overeating, or not exercising?

Annot	Line	Spkr	Conversation
. . .			
T	[44]	2:	Uh, I'm in college so, like, my drinking is pretty cheap.
	[45]	2:	Maybe like five bucks a week.
	[46]	1:	Oh, that's not bad.
	[47]	2:	[LAUGH] Yeah, it's pretty cheap.
	[48]	1:	Mhm.
S	[49]	1:	Wait, what college do you go to by the way?
	[50]	2:	University of Illinois.
	[51]	1:	Really, in Champagne?
	[52]	2:	Yeah. In Champagne.
	[53]	1:	Oh, wow.
	[54]	2:	And you live in New York?
	[55]	1:	Yeah.
	[56]	2:	Interesting.
	[57]	1:	Yeah.
M	[58]	1:	But - um - So anyways I guess we're off topic again [LAUGH].
	[59]	2:	[LAUGH] Yeah
	[60]	1:	Um- [LAUGH] um, what were the other things on the list?
	[61]	1:	Oh yeah, overeating.
T	[62]	1:	See, you know what I heard about, um, overeating is that -- or -- or just in general, like, you know, obesity and everything is that, um --
	[63]	1:	Right now smoking is the number one cause of death in the country.
	[64]	1:	But then pretty soon it's going at -- um, switch over to obesity.
	[65]	2:	Yeah. I've -- I've heard about that too.
. . .			

On-topic Keywords: **drinking**, **obesity**, **overeating**, **smoking**, **mortality**

References

[ALS05] Jeremy Ang, Yang Liu, and Elizabeth Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. Proceedings of ICASSP, 2005.

[BBL99] Doug Beeferman, Adam Berger, and John Lafferty. Statistical Models for Text Segmentation. *Machine Learning*, 1999.

[BBM04] Sugato Basu, A. Banjeree, and E. R. Mooney. Active semi-supervision for pairwise constrained clustering. Unknown, 2004.

[BC99] Timothy Bickmore and Justine Cassell. Small Talk and Conversational Storytelling In Embodied Conversational Interface Agents. 1999.

[BC00] Timothy Bickmore and Justine Cassell. How about this weather?: Social Dialogue with Embodied Conversational Agents. 2000.

[BMWK05] Rebecca Bates, Patrick Menning, Elizabeth Willingham, and Chad Kuyper. Meeting Acts: A Labeling System for Group Interaction in Meetings. August 2005.

[Bri92] Eric Brill. A simple rule-based part of speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, 1992.

[CB02] Justine Cassell and Timothy Bickmore. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and Adaptive Interfaces*, (12):1–44, 2002.

[Che88] Christine Cheepen. *The Predictability of Informal Conversation*. Pinter Publishers, London, 1988.

[Cho05] Freddy Choi. Advances in domain independent linear text segmentation. Proceedings of NAACL'00, Seattle, USA, April 2005.

[GJ02] Daniel Gildea and Daniel Jurafsky. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288, 2002.

[Hea97] M. A. Hearst. TextTiling: Segmenting Text into Multiparagraph Subtopics Passages. *Computational Linguistics*, 23(1):33–64, 1997.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. Proc. ECML, 1998.

[Kat96] S. M. Katz. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2(1):15–59, 1996.

[Lav75] J. Laver. *The organization of behavior in face-to-face interaction*, chapter Communicative functions of phatic communion, pages 215–238. Mouton, The Hague, 1975.

[Lav81] J. Laver. *Conversational routine*, chapter Linguistic Routines and politeness in greeting and parting, pages 289–304. Mouton, The Hague, 1981.

[LDC04] LDC. Fisher english training speech part 1, transcripts, 2004. http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T19.

[LG94] David D. Lewis and William A. Gale. A sequential algorithm for training text classifiers. Proceedings of SIGIR, 1994.

[Mit97] Tom Mitchell. *Machine Learning*. McGraw-Hill, Boston, 1997.

[MN98] Andrew Kachites McCallum and Kamal Nigam. Employing em and pool-based active learning for text classification. Proceedings of the 15th International Conference on Machine Learning, 1998.

[MS99] Christopher D. Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[Nig01] Kamal Paul Nigam. Using unlabeled data to improve text classification, 2001.

[NMTM99] Kamal Nigam, Andrew Mccallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 1999.

[Off03] DARPA Information Processing Technology Office. Effective, affordable, reusable speech-to-text (EARS), 2003. http://www.darpa.mil/ipto/programs/ears/.

[OSSB01] D. Oppermann, F. Schiel, S. Steininger, and N. Beringer. Off-talk - a problem for human-machine interaction, 2001.

[PH02] L. Pevzner and M. Hearst. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics*, 28(1):19–36, March 2002.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. Proceedings of EMNL, 2002.

[Rey98] Jeffrey C. Reynar. Topic segmentation: Algorithms and applications, 1998.

[Rey99] Jeffrey C. Reynar. Statistical models for topic segmentation. Proceedings of the 37th Annual Meeting of the ACL, 1999.

[RM01] Nicholas Roy and Andrew McCallum. Toward optimal active learning through sampling estimation of error reduction. Proceedings of the 18th International Conference on Machine Learning (ICML-01), 2001.

[San90] B. Santorini. Part-of-speech tagging guidelines for the penn treebank project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.

[Sch88] K. Schneider. *Small Talk: Analysing Phatic Discourse*. Hitzeroth, Marburg, Germany, 1988.

[Seb02] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[SM99] Sam Scott and Stan Matwin. Feature engineering for text classification. Proceedings of the 16th International Conference on Machine Learning, 1999.

[SRM04] Charles Sutton, Khashayar Rohanimanesh, and Andrew McCallum. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. Proceedings of the 21st international conference on Machine Learning, 2004.

[Sto04] Nicola Stokes. Applications of lexical cohesion analysis in the topic detection and tracking domain, 2004.

[TAJ04] Douglas P. Twitchell, Mark Adkins, and Jay F. Nunamaker Jr. Using speech act theory to model conversations for automated classification and retrieval. Proceedings of LAP, 2004.

[TK00] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. Proceedings of the 17th International Conference on Machine Learning (ICML-00), 2000.

[UI01] M. Utiyama and H. Isahara. A statistical model for domain - independent text segmentation. pages 491–498. Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics, 2001.

[VLSS05] Anand Venkataraman, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. Does active learning help automatic dialog act tagging in meeting data? Proceedings of Interspeech, 2005.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition.* Elsevier, 2005.

[WKNN97] V. Warnke, R. Kompe, H. Niemann, and E. Noth. Integrated dialog act segmentation and classification using prosodic features and language models. volume 1, pages 207–220. Proceedings of EUROSPEECH, 1997.

[Yan97] Yiming Yang. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1997.

[ZLSS05a] Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. A* based segmentation and classification of dialog acts in multiparty meetings. Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU), 2005.

[ZLSS05b] Matthias Zimmermann, Yang Liu, Elizabeth Shriberg, and Andreas Stolcke. Toward joint segmentation and classification of dialog acts in multiparty meetings. Proceedings of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2005.