Off-Topic Detection in Conversational Telephone Speech

Robin Stewart, Andrea Danyluk, and Yang Liu

ACTS Workshop, HLT-NAACL 2006

June 8, 2006

Introduction

- In the context of information retrieval of spoken documents, we assume for this project that users seek credible information about a specific topic.
- Some spoken utterances serve a different purpose: *"Nice weather we've been having."*
- Goal: automatically identify "irrelevant" utterances in the domain of telephone conversations.

Sample Conversation

Recorded telephone conversations with an assigned topic.

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

Linguistic Background

Two primary goals in conversation (Cheepen 1988):

- **transactional** goals, which focus on communicating useful information or getting a job done.
- **interactional** goals in which interpersonal motives such as social rank and trust are primary

Linguistic Background

Two primary goals in conversation (Cheepen 1988):

- **transactional** goals, which focus on communicating useful information or getting a job done.
- **interactional** goals in which interpersonal motives such as social rank and trust are primary

Approximate **transactional** vs. **interactional** with:

- **relevant** vs. **irrelevant** (to a task)
- **on-topic** vs. **off-topic**

Linguistic Background

Two primary goals in conversation (Cheepen 1988):

- **transactional** goals, which focus on communicating useful information or getting a job done.
- **interactional** goals in which interpersonal motives such as social rank and trust are primary

Approximate **transactional** vs. **interactional** with:

- **relevant** vs. **irrelevant** (to a task)
- **on-topic** vs. **off-topic**

Should be generalizable to other domains with a topic:

- broadcast debates, class lectures, meetings

Methodology

Empirical approach:

1. **Define** on- and off-topic.
2. Select **data**.
3. **Annotate** the data according to the definitions.
4. Generate **features** to describe each utterance.
5. Use **machine learning** algorithms to train classifiers on different feature sets.

Utterance-level classification.

Definitions

Classify **utterances** based on these definitions:

- **On-Topic**: the conversants are discussing something at least tangentially related to the assigned topic for the conversation.
- **Metaconversation**: 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.").
- **Small Talk**: includes 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 ("Oh, I live in a condo in Atlanta."), and weather ("I hear it's pretty hot down there...").

Data Selection

- Full data set had 5727 conversations.
- We randomly chose 4 conversations in each of 5 topics:
  - Computers in Education
  - Pets
  - Terrorism
  - Censorship
  - Bioterrorism
- This set of 20 conversations includes a total of **5070 utterances**.

Annotation

- Assign one of the labels (S, M, or T) to each utterance in a conversation.
- Each conversation annotated by 2-3 people
- Pairs of annotators agreed with each other on 86.1% of utterances.
- Need to deal with the 14% with mismatched labels:
  - On-Topic and Metaconversation "safer" than Small Talk
  - Only label Small Talk if all annotators agreed on it
  - On-Topic if *any* annotator thinks it's relevant.
- Result:
  - 17.8% Small Talk
  - 9.4% Metaconversation
  - 72.8% On-Topic

Creating Features

Each utterance is represented as a **feature vector** for the classifier.

Related research in the linguistics of conversational speech led us to hypothesize that certain features might be indicative of off-topic speech:

1. position in the conversation (Cheepen 1988),
2. the use of present-tense verbs (Cheepen 1988),
3. a lack of common helper words such as "it", "there", and forms of "to be" (Laver 1981).
   - *"Nice day."*

Features

- Position in the conversation
  - Represented by the line number (binned).
- Verb tense and parts of speech
  - We used Brill's tagger to automatically label the standard Penn part-of-speech tag for each word in the data set.
  - The features consist of the counts for each part-of-speech tag in a given utterance.
- Words
  - Bag-of-words model: counts for each word.
  - To choose which words to consider (limited memory), we used Lewis and Gale's (1994) **feature quality measure**.
    - Rationale: used for similarly short fragments of text.

Features: Other

- Utterance type (statement, question, or fragment).
- Utterance length (number of words in the utterance).
- Number of laughs in the utterance.

- Summary features for previous 5 and subsequent 5 utterances.

Notes about the features

- There is some overlap between features: The token "?" can be represented as:
  - A word (chosen by the feature quality measure)
  - A part-of-speech tag
  - Implicit in the utterance type (question)

- The conversation topic is *not* taken to be a feature
  - Looking for a more general characterization of on- and off-topic regions.
  - Topic information is not necessarily available.

Experimental Setup

- Chose the **SVM** algorithm because of its superior performance over the other ML techniques we tried (see paper).
- To test each feature set, we performed 4-fold cross-validation
  - Trained on 3 of the conversations in each topic (15 total).
  - Tested on the remaining 1 in each topic (5 total).
- We systematically varied the feature sets:
  - All features (for reference)
  - All of the features except one
  - One feature at a time
- Evaluation metrics:
  - Accuracy
  - Cohen's Kappa statistic

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

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

Implications for Linguistic Hypotheses

1. As expected, conversations in our data set have a predictable structure in that they routinely start with small talk.
   - A classifier with no information except line number labeled 17% of the small talk in the ten-minute conversations.

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

Implications for Linguistic Hypotheses

1. As expected, conversations in our data set have a predictable structure in that they routinely start with small talk.
   - A classifier with no information except line number labeled 17% of the small talk in the ten-minute conversations.
2. Contrary to our hypothesis, part-of-speech tags do not appear to contain useful information for distinguishing between utterance types.
   - 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.

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

Implications for Linguistic Hypotheses

1. As expected, conversations in our data set have a predictable structure in that they routinely start with small talk.
   - A classifier with no information except line number labeled 17% of the small talk in the ten-minute conversations.
2. Contrary to our hypothesis, part-of-speech tags do not appear to contain useful information for distinguishing between utterance types.
   - 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.
3. The types of words that proved useful for distinguishing amongst categories did not uphold the hypothesis that a lack of common helper words might be indicative of small talk.
   - Some of the words make intuitive sense as being important.
   - But overall they do not present a clear pattern.

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

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

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

Other Findings

- Utterance type, utterance length, and laughs are not very important
- The context of an utterance is important
  - Kappa statistic is twice as high when prev/next included

Other Findings

- Utterance type, utterance length, and laughs are not very important
- The context of an utterance is important
  - Kappa statistic is twice as high when prev/next included

More generally...

- Small Talk, Metaconversation, On-Topic are identifiable
- Words are the most crucial features
  - Highest accuracy and Kappa when used alone.

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

Other Findings

- Utterance type, utterance length, and laughs are not very important
- The context of an utterance is important
  - Kappa statistic is twice as high when prev/next included

More generally...

- Small Talk, Metaconversation, On-Topic are identifiable
- Words are the most crucial features
  - Highest accuracy and Kappa when used alone.
  - But words do "include" part-of-speech information.

Future Work

More candidate features:

- Parse structure
- Timing and pause duration
- Prosodic information

Improve the detection system:

- Other approaches to classification and segmentation.
- More data.
- Speech-recognized transcriptions.

Broaden the scope of analysis to new genres:

- Broadcast news, class lectures, meetings

Acknowledgements

- Advice:
  - Mary Harper
  - Brian Roark
  - Jeremy Kahn
  - Rebecca Bates
  - Joe Cruz
- Student annotators:
  - Nick Anderson
  - Mary Beth Anzovino
  - Sara Beach
  - Jessica Chung
  - Jonathan Dowse
  - Kathryn Fromson
  - Caroline Goodbody
  - Ikem Joseph
  - Katie Lewkowicz
  - Lisa Lindeke
  - Myron Minn-Thu-Aye
  - Kenny Yim

Questions