Speaker Identification from Transcripts of TV Series

He He
University of Maryland College Park
hhe@cs.umd.edu

Ran Liu
University of Maryland College Park
ranliu@cs.umd.edu

Abstract

In this paper, we experiment with speaker identification with textual information. We not only take into account the features of the speech of certain character, but also included contextual information by analyzing the other participants in the same conversation.

1 Introduction

Given the transcript of a conversation in some context, humans are often able to identify speakers whom they are familiar with based on the syntactical style, the word usage, the topical characteristics of the utterance as well as the contextual information. In this project, our goal is to build a similar system that accepts a set of possible speakers and a transcript of segmented conversations as input and outputs the speaker for each utterance. This can also be used for labeling meeting notes and leverage audio-based speaker identification system.

We explored the potential of using only part-of-speech (POS) and word usage features for this task, with the hope to model individual's speaking style in a certain conversation. In addition, considering that a person may change his/her style depending on whom he/she is talking to, we designed features that including the contextual information so as to maximally model the real situation.

The paper is organized as follows: in Section 2, we present related work; in Section 3, we describe the dataset we used; in Section 3, we show the features we selected and the classification methods used in the learning part; in Section 4, we report our results and give some intuitive interpretation.

2 Related Work

The problem of speaker identification (diarization) has been extensively studied in the field of speech processing using audio features. However, to the best of our knowledge there are only a limited research on speaker identification using only texutal data. (Chaudhuri and Raj, 2011) used the Author-Recipient Topic Model (McCallum et al., 2007) to first identify a subset of speakers who are participating the conversation and then applied an exponential state Hidden Markon Model to model the speaking sequence. (Gillick, 2007) attempted to use only word-level information (bag of bigrams) with feature selection to predict speaker demographics from phone conversation transcripts. Recently, there is also growing interest in authorship identification in short text such as SMS (Mohan et al., 2011) and tweets (Sousa-Silva et al., 2011). These approaches extract sylistic features based on N-grams to recognize author of the short text.

In this project, besides experimenting with different set of idiosyncratic features, we try to take the contextual information in modelling the sequence in a conversation. More specifically, we include the speakers' possible addressees in the feature space.

3 Data

The data we used in the this paper is the scripts for the TV series The Big Bang Theory. For each line in the data file are formatted as speaker name, followed by the line for that character. The conversations are clustered and separated by the scene. In the original data set, each scene has a description of the

location, and some of the action descriptions for the actors. We removed these descriptions for the convenience of the experiment. But we keep the scene information to group the speeches by conversation.

We used the first two season of *T*he Big Bang Theory in our experiment. There are total 4137 iterations in season 1, and 5253 in season 2. We processed the original transcripts to only keep the name of the five main characters: Leonard, Sheldon, Howard, Raj and Penny. To keep the conversation information, we keep the speech of other characters but replaced the name as OTHER. In the experiment, we use the Big Bang Theory season 2 as our training set and season 1 as our testing data.

A sample data is shown below:

*S*cene: A corridor at a sperm bank.

Sheldon: So if a photon is directed through a plane with two slits in it and either slit is observed it will not go through both slits. If its unobserved it will, however, if its observed after its left the plane but before it hits its target, it will not have gone through both slits.

Leonard: Agreed, whats your point?

Sheldon: Theres no point, I just think its a good idea for a tee-shirt.

4 Method

4.1 Feature

Previous work has shown that the frequency distribution of part-of-speech (POS) tags may indicate text genre and can be helpful in text classification (Ott et al., 2011). In our case, it is also reasonble to use POS tags counts in a person's utterance. For example, some people tend to use many adjectives to describe one thing and geeky people may use more technical words that results in higher frequency of NN/NNP unigram and bigram tags. Therefore, we choose unigram and bigram POS tags as one set of features to experiment with.

In a topic modeling perspective, it is awared that although people use the same set of common words in daily conversations, each individual may have a subset of idosyncratic volcabulary. In our dataset, for instance, Howard talks a lot about women and loves to brag about his Master degree from MIT, thus the word "women" and "MIT" can be a strong indicator of the label "Howard". We include both unigrams and bigrams in the word feature, but only maintain those that occurs more than 5 times in the training set. We also keep a small amount of stopwords such as "well", "uh", "hmm" and so on, since they can reflect the speaker's speaking style to some extent. For example, when people are uncertain or conservative about the topic, they are likely to use these words so as to have more time for thinking.

Apart from POS-based and word-based features, we include punctuations, sentence length and average word length as well. However, based on our limited experiment, they degraded the system's performance.

Intuitively, it is known that people may change their speaking styple slightly when talking to different people and reponding to different topics. For example, Sheldon will always argue with his friends, but he will seldom argue or discuss with his mom. To include such contextual information, we concatenate the feature of an utterance with those of its previous and subsequent turns.

4.2 Model

4.2.1 Multiclass Classification

The speaker identification can be framed in a multiclass classification setting, where we have 6 classes: {Sheldon, Leonard, Howard, Penny, Raj, Other}. However, it is notices that in our case, the class distribution is imbalance, with a large amount of data for "Sheldon" and "Leonard", while very limited data for "Raj". We used different combinations of the feature sets described above to train a *M*aximum Entropy classifier.

In MaxEnt model, we estimate the conditional distribution of the class label y given the observation $\boldsymbol{x}$ using a log linear model.

$$p(y|\boldsymbol{x}) = \frac{1}{Z} \exp \sum_i w_i f_i \qquad (1)$$

Here, Z is a normalization factor that enforces the exponential to be a true probability; f_i denotes the features and w_i is the weight for the corresponding feature. Normally, at test time, we always predict whichever class that has the highest probability. However, our problem is indeed a structured prediction problem while at training time, each example is treated independently. To enforce the requirement

that speakers speaks in turn, which means there is no consecutive examples that have the same label, we predict the class that has the highest probability *a*nd is different from the previous class in a conversation.

4.2.2 Structured Prediction

In multiclass classification, we tried a tuple of concatenated features of three consecutive utterance to incorporating the information about the current speaker's addressees. Although the model does not have any structural knowledge, we encode it in the feature space in a naive manner. Another option is to just include the class label of the utterance's "neighbors" and train the classifier using stacking as in collective classification. Howevev, we believe that only the class label is not sufficient to provide the contextual information. Imagining in daily life, even if we are talking to the same person, we could change our speaking style according to the other's response/register and the current topic.

To further model the sequence of speaker turns in a conversation, we use a linear-chained Conditional Random Field (CRF). Since it can incorporate much richer features than a Hidden Markov Model (HMM) by directly modelling the conditional distribution insteand of the joint distribution to avoid the independence assumption of different features. Formally, we have

$$p(y|\boldsymbol{x}) = \frac{1}{Z} \exp\left\{\sum_i \lambda_k f_k(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)\right\} \quad (2)$$

Here $f_k(\boldsymbol{x}_{t-1}, \boldsymbol{x}_t)$ is the feature for both the current and the previous utterances.

5 Experiment

We used data from *T*he Big Bang Theory as described in Section 2. Data in the first season is used as testing set and data in the second season is used as training set. For the MaxEnt model, we used the MEGA Model Optimization Package (Daumé, 2011).

5.1 Result

The results are shown in Table 1. In this paper we explore two types of features: words and the Part Of Speech (POS) tag of the words. We ran the experiment on the data set. For each type of feature, we are using both of the unigram and bigram as the features. For the words feature, the result accuracy is 0.3442, slightly better than POS tag, which is 0.3200. To improve the result, we tried to use more sophisticated features. We removed the stop words from the conversations, hoping by remove the non-meaningful words from the feature, the result would improvement. But the result accuracy turned out opposed to what we hoped. We then tried to map the POS tags to the universal tags, but the result accuracy reduced again.

Type of Features	**Accuracy**
word	0.3442
word without stop words	0.3343
POS	0.3200
universal POS	0.3110
characters	0.3346

Table 1: Result of single type of features (each of experiment include the unigram and bigram of the corresponding feature type.

We then performed experiments using the combination of the features. The results are illustrated in Table 2. The most promising combination is word and POS tag, also include both unigram and bigram. The combination of words without stop words and POS tags also shown improvement than single feature. Almost all the combinations provide better results than single feature, except the combination of universal tag and word. Given the fact that universal tag does not perform well itself, the result is kind of expected.

We observe the combinations are slightly better than single characters feature. It might due to the features are not enough. The next step is to explore more features using features engineering techniques. But due to the time limitations, we did not address this issue.

6 Conclusion and Future Work

In this paper, we present both syntactical and topical features to predict the speaker from the transcripts of his/her utterance. Our experments show that textual information only can indicate speakers' speak-

Type of Features	Accuracy
POS + word	0.3469
UPOS + word	0.2890
POS + word(no stop words)	0.3445
POS + word(no stop words) + word length	0.3333
POS + word + word length	0.3345

Table 2: Result of combination of features (each of experiment include the unigram and bigram of the corresponding feature type.

ing style to some extent. We also showed that using structured prediction to incorporate contextual information improves the result.

In the future, we are very interested to use CRF to model the conversation sequence and use some feature selection criteria to reduce the feature space dimension.

References

Hal Daumé 2004. *Notes on CG and LM-BFGS Optimization of Logistic Regression*.

Myle Ott, Yejin Choi, Claire Cardie and Jeffrey T. Hancock 2011. *Finding Deceptive Opinion Spam by Any Stretch of the Imagination*. In *P*roceedings of ACL.

Ashwin Mohan, Ibrahim M. Baggili and Marcus K. Rogers 2011. *Authorship attribution of SMS messages using an N-grams approach*. *C*ERIAS Tech Reports 2010-11.

Rui Sousa-Silva, Gustavo Laboreiro, Luís Sarmento, Tim Grant, Eugénio C. Oliveira and Belinda Maia. 2011. *'twazn me!!! ;)' Automatic Authorship Analysis of Micro-Blogging Messages*. In *P*roceedings of NLDB.

Dan Gillick. 2010. *Can Conversational Word Usage Be Used to Predict Speaker Demographics?*. In *P*roceedings of Interspeech.

Andrew McCallum, Xuerui Wang and Andres Corrada-Emmanuel. 2007. *Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email.*. In *J*ournal of Artificial Intelligience Research.

Sourish Chaudhuri and Bhiksha Raj. 2011. *A Comparison of Laten Variable Models for Conversation Analysis*. In *P*roceedings of SIGDIAL.