Experimental Design

We evaluated our approach on two datasets: the Switchboard corpus (Godfrey, Holliman, & McDaniel 1992), a benchmark dataset for topic spotting and speech recognition, and the Google Answers dataset[5], a collection of short questions cataloged by topics, which we extracted from the web.

The Google Answers dataset is a collection of 8,850 questions pertaining to 9 top-level categories extracted from the web (around 1000 questions per category). The list of the categories is given in Table 3. Before Google Answers was discontinued, Google required a payment for answering a question which was answered by an expert in Google Answers. Therefore, the questions in this dataset are typically directed to experts, and require a lot of prior knowledge to be correctly categorized. Below are some examples of the questions in the Arts & Entertainment category:

1. *In 1998, Henry Rollins did a spoken word engagement gig in/near Venice beach... i'd like to know the date.*
2. *Please provide general information including best photos of beautiful "antigua town" of Guatemala country.*
3. *Looking for Boy Goergoe manager's phone number.*

The Switchboard corpus is a multi-speaker corpus of conversational speech with about 2500 conversations by 500 speakers from around the US. These conversations are transcribed by speaker terms and span 70 topics (like camping, taxes and recycling). A sample of the transcribed data is given below:

A.5: Uh, do you ha-, are you a musician yourself?
B.6: Uh, well, I sing.
A.7: Uh-huh.
B.8: I dont play an instrument.
A.9: Uh-huh.
Where, do you sing in, in a choir or a choral group?
B.10: Oh, not right now.

The Switchboard dataset was previously used for classifying transcripts in (Myers *et al.* 2000). Following (Myers *et al.* 2000), we manually map selected topics to ten categories as shown in Table 3. We consider the task of classifying each individual speaker turn to one of 10 categories. Our 10 category corpus has 46,000 utterances, with an average of 71.7 speaker turns per conversation. Since many of speaker turns do not carry meaningful information, We filter out stop words and sentences that contain less than 10 words to get 6840 sentences.

We use SVM with a linear kernel as our classification technique [6]. We report averaged results over five fold experiments using 20

Feature Selection

On the Google Answers dataset, the ESA feature expansion algorithm on the Wikipedia dataset generates 98006 features. Yahoo PAF generates 822 features, Open Directory Project generates 494 features, and clustering method

[5]Google Answers *http://answers.google.com/answers/*.

[6]We use the SVM Multi-Class implementation of (Tsochantaridis *et al.* 2004; Crammer & Singer 2001)

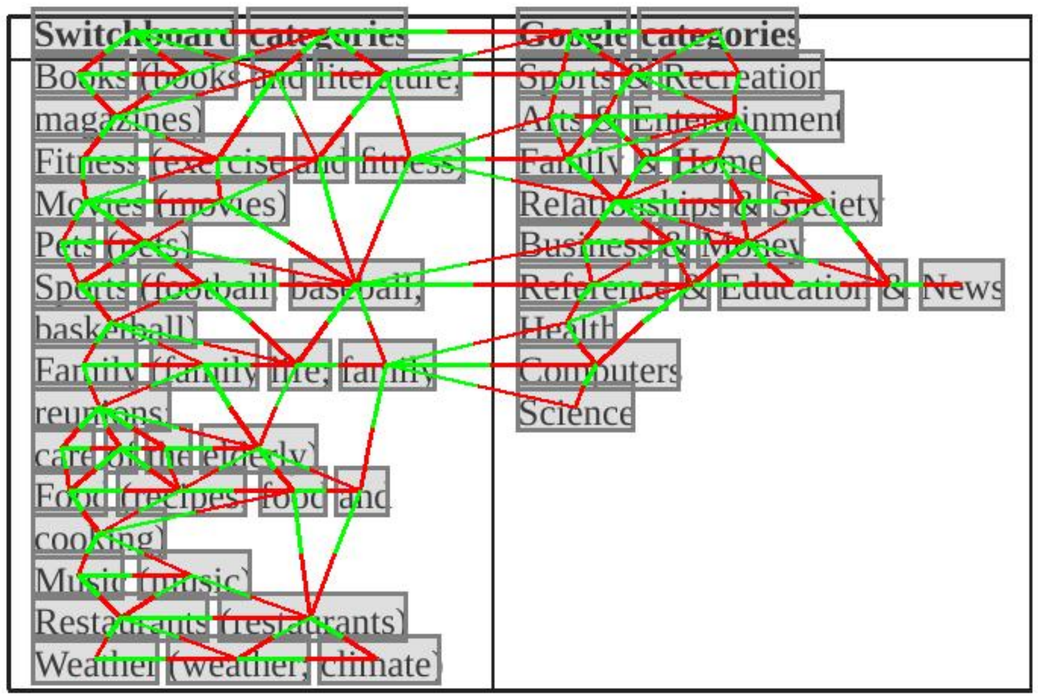

Table 3: Switchboard and Google categories

generate as many features as the number of clusters. With 11430 BOW features this gives a total of 112056 features. Clearly, to make the classifier more efficient and to avoid over-fitting, feature selection must be performed (Yang & Pedersen 1997). The basic feature classification algorithm is as follows. For each class c we compute the expected mutual information of the feature f and the class c. Mutual Information (MI) is zero if the feature distribution in the collection is same as in class and is maximum when the feature is a perfect indicator of the class. In the information theoretic sense, MI measures how much information the presence/absence of the feature f contributes to making the correct classification decision on c. Formally :

$$MI(f;c) = \sum P(F,C) log \frac{P(F,C)}{P(F)P(C)}$$

where F is a binary random variable that takes the value 1 if sample x contains the feature f and C is a binary random variable that takes the value 1 if x belongs to category c. We use maximum likelihood to estimate these probabilities (Manning, Raghavan, & Schtze 2008).

Table 4 shows top 10 features with highest MI scores for the selected classes in the Google Answers dataset. Surprisingly, using only these 10 features per class, it is possible to achieve 63.39% accuracy for the Google dataset, which is better than using bag-of-words (BOW) approach with best performance at 46.55% with 200 features/class. ESA features shown in green are quite good and intuitive. PAF features generated from Yahoo dataset shown in red and the PAF features generated from ODP features shown in blue. Most interestingly, some of the features do not have an obvious correlation to the category. For example, the Yahoo subcategory *'Entertainment&Music- Jokes'* is not obviously highly correlated to *'Sports and Recreation'* category in Google Answers. However, its top-20 TF-IDF words summarization empirically explains the target category well. We note that MI is a greedy method because in the Health category *cancer* is selected as a feature even though it is highly correlated with *prostate cancer* which is previously selected as a feature and is therefore redundant.