Leveraging Locality for Topic Identification of Conversational Speech

Jonathan Wintrode

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, Maryland
jcwintre@cs.jhu.edu

Abstract

We evaluate the limitations of the bag-of-words assumption for topic identification of conversational discourse by examining whether topic-dependent word occurrence statistics are also position-independent. We demonstrate where the assumption is violated in conversational speech corpora and show how the relevance of words to the classification task decreases over the length of the document. We seek to improve topic identification by modeling this topic drift phenomenon and weight word counts according to a decay function over the length of the document. By applying a global decay rate for all words we observe a reduction in error rates of 23-47% relative on conversational corpora. Furthermore, we apply a minimum classification error (MCE) training procedure to learn per-word decay rates, and reduce error rates by up to an additional 27%.

Index Terms: text categorization, bag-of-words assumption, discriminative training, Naive Bayes, topic identification

1. Introduction

The bag-of-words assumption is widely used in text categorization and information retrieval, and widely assumed to be incorrect. Alternative approaches have been tried with varying success, but the simplicity and computational tractability of the bag-of-words assumption is compelling. The bag-of-words assumption states that the content of a document can be captured by a function of the frequency of the words it contains, independent of position. We can express arbitrarily large documents by a sparse vector of word counts. We will show that this position-independence claim is specifically problematic for conversational discourse. By incorporating position information into a Naive Bayes classification model, we can improve performance on text categorization tasks in this domain.

We propose a simple framework to model the phenomenon of topic drift, in which the contribution of individual occurrences to the effective word frequency decreases over the course of the document. We begin by assuming a global decay rate for all words. We then learn optimal decay rates for each word using a minimum classification error (MCE) training framework. We show that for some conversational corpora we are able to cut identification errors in half.

1.1. Related Work

We focus on the bag-of-words assumption in the context of a Naive Bayes classifier. The combination of the two has been extensively studied in the machine learning and information retrieval communities. A general treatment of the Naive Bayes modeling assumptions, including the bag-of-words assumption, is given by Lewis in [1] and McCallum et al. in [2]. The combination of the bag-of-words word frequency assumption and Naive Bayes arises in the multinomial event model, described in detail in [2].

A simple extension of the unigram bag-of-words model is the n-gram model, which captures local sequence information of up to n words, but still assumes that n-gram frequencies are independent of position. In the context of text categorization, both Boulis [3] and Tan [4] found that a mixture of unigrams and bigrams outperformed all of one or the other. Alternatively there have been attempts to add non-word features to the document representation (e.g. [5]).

Other work has focused on feature weighting or feature selection ([6], [7], [8], and [9]). Work by Hazen and Margolis [8] and Boulis [3] are among the few efforts to focus on speech corpora rather than exclusively text corpora (Fisher English, in both cases). In general, improvements have been demonstrated over the basic bag-of-unigrams model.

Some related work has been done in the field of summarization, relating topic position and discourse structure [10]. Referring much earlier work by Edmundson looking at the discourse structure of text, the observation is made that the topical content is more prevalent at the beginning and end of a document [11]. However, given the conversational corpora at our disposal, we observe the effect only at the beginning, not the end.

Our model of topic drift is an example of a technique that modifies the per-document term frequency (TF) for text categorization. A collection of such techniques were proposed by Xue and Zhou [12] who aim to replace TF in traditional vector-space TF-IDF weighting schemes with a number of alternative relevance weightings, some based on within-document position and improving classification performance on the Reuters-21578 data. Rennie et al. [13] also propose a number of re-weighting and normalization techniques, but not specifically tied to position. Work by Lebanon et al. [14] presents a formal generative model that incorporates position information through kernel smoothing of frequency estimates, and demonstrates improvement on the top 10 topics in the Reuters-21578 corpus.

2. Experiment Details

Topic identification (or text categorization) techniques can be used for a number of related tasks. These are:

- Identification: Assign a specific topic label to a document.
- Detection: For each topic label, give a score to indicate whether the topic occurred within the document.
- Retrieval: Return to the user a ranked list of documents relevant to a given topic label.

We focus on the performance of these related tasks on conversational or informal discourse. Examples of such data in-
clude online videos, meeting speech, or voicemail, in contrast to broadcast, lecture or other formal speech. We expect informal communications to differ from formal ones in vocabulary, linguistic content, discourse structure, and, as will the focus of this work, topic structure.

In order to evaluate classification performance, we use the topic-labeled Fisher English [15] and Spanish [16] conversational speech transcripts, collected by LDC. Participants in the Fisher collection paradigm were prompted to speak on a particular topic without any other constraints. Thus document-level topic annotations are available implicitly.

For English, we use the same split as in previous work ([8], [17], [18]), with 1374 conversations for training and 686 for testing, distributed over 40 topic labels. We induced a randomly sampled 80/20 train/test split on the Fisher Spanish transcripts, distributed over 25 different topic labels.

2.1. Metrics
A number of metrics are typically used to evaluate classification tasks. Text categorization work has reported identification (ID) error, micro and macro-averaged F-score. These measures require the system to make a hard labeling decision for each document, as is necessary for certain applications such as document routing. We also report the area under the precision-recall curve (AUC) averaged over all topics, and the equal error rate (EER) point, averaged over all topics. These two metrics are independent of a specific threshold or operating point and are only sensitive to the ordering of results. We are interested both in metrics for the ID (labeling) task (Error, F-score) and for the retrieval/detection task (AUC, EER).

3. Naive Bayes Classification

The Naive Bayes assumption states that the words of a document are conditionally independent given the topic. So given a document \( D \) that is \( |D| \) words long and a topic \( t \), we can compute its likelihood:

\[
P(D|t) = \prod_{i=1}^{l(D)} P(w_i|t) \tag{1}
\]

Training of a Naive Bayes classifier is done simply by estimating these conditional probability distributions from training data.

We can perform the ID task using a log-likelihood ratio (LLR) test, as in Equation 2, or we can treat the LLR test as a scoring function for detection or retrieval (Equation 3).

\[
i = \arg \max_i \log \left( \frac{P(D|t_i)}{P(D|\bar{t}_i)} \right) \tag{2}
\]

\[
S(D|t) = \log \left( \frac{P(D|t)}{P(D|\bar{t})} \right) \tag{3}
\]

3.1. Bag-of-Words Formulation

Typically, the Naive Bayes assumption is paired with the bag-of-words assumption of position independence. We can replace same-word occurrences from Equation 1 with counts \( c_w \) to obtain the typical Naive Bayes formulation for \( P(D|t) \):

\[
\log \left( P(D|t) \right) = \sum_{w \in D} c_w \cdot \log \left( P(w|t) \right) \tag{4}
\]

Table 1 shows baseline results using the typical approach. While there are better results on the Reuters data using other classifiers, these are consistent with other published results using Naive Bayes [13]. As discussed previously, this simple bag-of-words assumption has proven difficult to improve on. However, using the Fisher corpora, we demonstrate that for conversational discourse, the location independence assumption is flawed and can be relaxed with positive performance gains.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Error</th>
<th>AUC</th>
<th>EER</th>
<th>MicroF1</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters</td>
<td>20.1%</td>
<td>0.531</td>
<td>0.160</td>
<td>0.701</td>
<td>0.308</td>
</tr>
<tr>
<td>Fisher Eng</td>
<td>11.2%</td>
<td>0.879</td>
<td>0.048</td>
<td>0.888</td>
<td>0.871</td>
</tr>
<tr>
<td>Fisher Spa</td>
<td>25.0%</td>
<td>0.823</td>
<td>0.122</td>
<td>0.750</td>
<td>0.743</td>
</tr>
</tbody>
</table>

Table 1: Baseline Naive Bayes results.

3.2. Topic Drift

For topic identification, the location independence assumption implies that the topic label for a document applies uniformly across all words. We test this assumption empirically by building separate Naive-Bayes models on one quartile from each conversation at a time. For example, we consider only the first 25% of each conversation when training and testing the Naive Bayes classifier, then the second, and so on (Table 2). If document content were independent of position, then we would expect error rates to be roughly the same no matter what quartile we trained and tested on, as is the case for Reuters.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>0-25%</th>
<th>25-50%</th>
<th>50-75%</th>
<th>75-100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reuters</td>
<td>23.5%</td>
<td>21.9%</td>
<td>21.7%</td>
<td>21.2%</td>
</tr>
<tr>
<td>Fisher Eng</td>
<td>8.9%</td>
<td>24.3%</td>
<td>38.5%</td>
<td>43.6%</td>
</tr>
<tr>
<td>Fisher Spa</td>
<td>25.6%</td>
<td>34.7%</td>
<td>42.1%</td>
<td>48.3%</td>
</tr>
</tbody>
</table>

Table 2: Topic ID error observed when training and testing on data by quartile.

For the Fisher corpora performance is significantly better when using only the first 25% of each document. Performance degrades the further into the conversation we get. This phenomenon, which we’ll refer to as topic drift, indicates to us a quantifiable difference between the formal (Reuters) and informal (Fisher) discourse.

4. Static Drift Model

Given these observations, we propose a model in which the count for each word is weighted relative to its position in the document. To match the drift over time, we propose a set of decay functions \( d(p, \lambda) \) that decrease over the range \([0, 1]\) in order to model decreasing word relevance the further we go into the conversation.

\[
d_{\text{exp}}(p, \lambda) = \exp(-\lambda \cdot p) \tag{5}
\]

\[
d_{\text{gauss}}(p, \lambda) = \exp\left(-\frac{\lambda^2 \cdot p^2}{2}\right) \tag{6}
\]

\[
d_{\text{lin}}(p, \lambda) = \begin{cases} 1 - \lambda \cdot p : & p \in [0, 1] \\ 0 : & p \notin [0, 1] \end{cases} \tag{7}
\]

Equation 5 is an exponential decay, Equation 6 is an unnormalized Gaussian curve, and Equation 7 is a linear decay. Using
any of these we can state our decay-weighted count estimate as follows:

\[ c_w = \sum_{i=1}^{[D]} d_i \frac{i}{[D]} \cdot I_w(w_i) \]  

(8)

Where \( I_w(w_i) \) is an indicator function whose value is 1 where \( w_i = w \) and 0 otherwise.

### 4.1. Results

Using our new definition for \( c_w \), weighted by position, we build classifiers using different values of \( \lambda \) on the Reuters and Fisher datasets. We have defined our decay functions such that an increase in \( \lambda \) implies a higher rate of decay and relatively more emphasis on the words at the beginning of the document. The results, shown in Figure 1, present identification error at values ranging from 0 (the baseline where all weights are 1) to 5.

For Reuters, the topic drift model is somewhat detrimental to performance, as we did not expect any measurable topic drift to have occurred in this genre. The trend is consistent using metrics other than error. For the conversational data, Fisher English and Spanish, we observe up to a 5% absolute reduction in classification error and 23% to 47% relative reductions.

#### Figure 1: Classification error vs. \( \lambda \)

![Classification error vs. \( \lambda \)](image)

5. **Discriminative Model**

We extend the static drift model by allowing the relevance of individual words to decay at different rates. Instead of a global \( \lambda \) value for all words, as in Equation 8, we calculate an optimal \( \lambda_w \) for each word. With this modification, our estimate for \( c_w \) becomes:

\[ c_w = \sum_{i=1}^{[D]} d_i \frac{i}{[D]} \cdot I_w(w_i) \]

(9)

We propose to estimate \( \lambda_w \) using a minimum classification error (MCE) training [19]. The derivation of the gradient-descent update equations follows that given by [8]. The major difference between derivations is that the parameter \( \lambda_w \) we wish to optimize occurs inside the decay function, so we are obliged to take the partial derivative of \( d(p, \lambda_w) \) as well.

Modifying the scoring function by combining Equation 9 with Equation 4, we obtain \( S(t|D) \) where the per-word contribution is a weighted sum over each position, rather than a single term.

\[ S(t|D) = \sum_{i=1}^{[D]} d_i \frac{i}{[D]} \lambda_w \cdot \log \left( \frac{P(w_i|t)}{P(w_i|t_j)} \right) \]

(10)

We use the same misclassification measure as in [8], \( M(D) \), and loss function \( l(D) \), which maps \( M(D) \) to a \([0, 1]\) range. Here \( t_C \) is the correct topic label for \( D \) and \( t_j \) is the incorrect topic with the highest score. For notational simplicity, we also define \( M(w) \) as the word specific component of \( M(D) \).

\[ M(D) = S(t_C|D) - S(t_j|D) \]

(11)

\[ l(D) = \frac{1}{1 + e^{-\beta M(D)}} \]

(12)

\[ M(w) = \log \left( \frac{P(w|t_C)}{P(w|t_C)} \right) - \log \left( \frac{P(w|t_C)}{P(w|t_j)} \right) \]

(13)

We now compute the partial derivative and update equations for gradient-descent optimization of \( M(D) \), which contain our decay function \( d(p, \lambda_w) \).

\[ \frac{\partial l(D)}{\partial \lambda_w} = \beta l(D)(1 - l(D))M(w) \left( \sum_{i=1}^{[D]} \frac{\partial d_i}{\partial \lambda_w} \right) \]

(14)

\[ \lambda_w' = \lambda_w - \epsilon \frac{1}{N} \sum_{j=1}^{N} \frac{\partial l(D_j)}{\partial \lambda_w} \]

(15)

Given the computational cost of performing the gradient descent, we evaluate the MCE training using only the exponential and Gaussian decay functions, which performed better in our static tests. The partial derivatives for \( d(p, \lambda_w) \) are given as follows:

\[ \frac{\partial d_{\exp}}{\partial \lambda_w} = -p \cdot \exp(-\lambda_w \cdot p) \]

(16)

\[ \frac{\partial d_{\gauss}}{\partial \lambda_w} = -p^2 \cdot \lambda_w \cdot \exp \left( -\frac{\lambda_w^2 \cdot p^2}{2} \right) \]

(17)

It remains to choose appropriate \( \epsilon \) (rate) and \( \beta \) (scale) parameters, as well as stopping criteria. In our experiments we use 5-fold cross-validation to compute the training loss. We found empirically that for the English data \( \epsilon = 10 \) and \( \beta = 0.01 \)
achieved the best results, whereas for Spanish $\epsilon = 100$ and $\beta = 0.1$ were best. We recognize the need for more extensive testing of hyper-parameters, lack additional held-out development data to perform those experiments. Results are reported on the held-out test data.

5.1. Results

Using the discriminative training approach, the best configurations were able to achieve an additional 7% (English) and 27% (Spanish) reduction in error, relative to the best static decay weights. In English, the lowest ID error achieved was 5.5%, a 51% relative reduction over the baseline. In Spanish the lowest error was 14.2%, a 43% relative (11% absolute) reduction.

We initialized $\lambda_w = 4$ for all words, starting with the best-performing static $\lambda$ from the previous section’s results. We achieved proportional gains above static decay weights using other initial values, but they did not always converge to the absolute best. This set of initial conditions gave the overall best results with the MCE training.

With these settings, we stopped observing gains after about 2000 iterations of MCE training. In some cases, over-fitting began to occur after this point and performance on the test data began to decrease. We observed similar gains in the other metrics as well, and the best results are listed in Table 3. It is encouraging that we were able to take our best static results, begin MCE training at that point and see additional gains beyond that point. Also, it is not surprising that the best static model yielded a good set of initial conditions for the discriminative training.

The distribution of the weights after 2000 iterations is still heavily centered around the initial $\lambda = 4$. For Spanish, the largest weights range well above 4, up to 10, whereas for English, the largest decay rate is 5.4. Table 4 shows the words with the highest learned decay rate in Fisher Spanish. The mixture of stopwords and place names at first seems surprising, but place names were used by participants during introductions, were unrelated to the actual topic, and were given a high decay rate and effectively ignored by the algorithm. Traditional frequency methods of stopword detection would have a hard time flagging the unique place names as stopwords.

6. Conclusions

What we have proposed is a modification of the bag-of-word assumption for topic classification in order to model conversation participants drifting off their main topic. We model topic drift with a decay function applied to the computation of term frequency, and then proceed to build a traditional bag-of-words classifier. This simple model, when combined with MCE training, is able to learn per-word decay rates, gives us a solution that reduces the topic ID error rate on Fisher conversational corpora by up to 51% relative an unweighted Naive Bayes baseline.

It is worth observing that the decay-models improved performance in both English and Spanish conversational corpora, but not on the text (Reuters) classification task, where we do not expect the phenomenon to occur. This suggests a genre-specific phenomenon, not a language specific one.

We do recognize that the topic drift in the Fisher corpora is due in part to the manner in which participants were assigned a topic. Presumably they begin talking about the prescribed topic, because they feel obliged to, then drift off-topic. In the future, we would like to explore if a decay-based model could be adapted to additional conversational scenarios or more complex discourse structures, without assuming the topic of interest occurs primarily at the start of the conversation. The amount of labeled data for Fisher makes it a useful starting point, but a starting point nonetheless.

7. Acknowledgments

The author wishes to thank Wade Shen for suggesting the document-initial topic bias of the Fisher English corpus, as well as TJ Hazen and Anna Margolis for the clarity of their 2008 paper deriving MCE training, making it simple and straightforward to adapt the technique to new problems.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>$\lambda$</th>
<th>Decay</th>
<th>Iterations</th>
<th>Error</th>
<th>EER</th>
<th>AUC</th>
<th>MicroF1</th>
<th>MacroF1</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>11.2%</td>
<td>4.8%</td>
<td>0.879</td>
<td>0.888</td>
<td>0.871</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{gauss}$</td>
<td>0</td>
<td>7.3%</td>
<td>2.7%</td>
<td>0.942</td>
<td>0.927</td>
<td>0.908</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{exp}$</td>
<td>2000</td>
<td>5.7%</td>
<td>2.6%</td>
<td>0.944</td>
<td>0.943</td>
<td>0.923</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{exp}$</td>
<td>2000</td>
<td>6.0%</td>
<td>2.7%</td>
<td>0.945</td>
<td>0.940</td>
<td>0.927</td>
</tr>
<tr>
<td>Spanish</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>25.0%</td>
<td>12.2%</td>
<td>0.823</td>
<td>0.750</td>
<td>0.743</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{gauss}$</td>
<td>0</td>
<td>19.9%</td>
<td>9.7%</td>
<td>0.840</td>
<td>0.801</td>
<td>0.782</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{exp}$</td>
<td>2000</td>
<td>15.3%</td>
<td>7.2%</td>
<td>0.872</td>
<td>0.847</td>
<td>0.823</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$d_{exp}$</td>
<td>2000</td>
<td>19.3%</td>
<td>10.2%</td>
<td>0.845</td>
<td>0.807</td>
<td>0.791</td>
</tr>
</tbody>
</table>

Table 3: Best results for each method

<table>
<thead>
<tr>
<th>Word (Decay Rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>hm 10.441</td>
</tr>
<tr>
<td>guatemala 10.189</td>
</tr>
<tr>
<td>asunción 9.950</td>
</tr>
<tr>
<td>acá 9.852</td>
</tr>
<tr>
<td>oh 9.090</td>
</tr>
<tr>
<td>sí 9.008</td>
</tr>
<tr>
<td>uh 8.859</td>
</tr>
<tr>
<td>méjico 8.657</td>
</tr>
<tr>
<td>ajá 8.518</td>
</tr>
<tr>
<td>bonito 8.397</td>
</tr>
</tbody>
</table>

Table 4: Highest decay-rate words after 2000 MCE training iterations on Fisher Spanish
8. References


