IMPROVING LATENT SEMANTIC INDEXING BASED CLASSIFIER WITH INFORMATION GAIN

Li Li, Wu Chou
Avaya Labs Research
233 Mountain Airy Road
Basking Ridge, New Jersey 07920, USA
lli5@research.avayalabs.com
wuchou@research.avayalabs.com

ABSTRACT
In this paper, we describe an approach of using a discriminative term selection process based on information grain (IG) to improve the performance of the latent semantic indexing (LSI). The discriminative power of the term is measured by entropy variations averaged over all categories conditioned upon whether the term is present or absent. The proposed approach is applied to the task of natural language call routing (NLCR), where natural language based classifiers are used to route calls to desired destinations. Various experimental studies are performed. Significant performance gains of 27% on precision and 26.5% on recall are observed. Most importantly, the proposed approach is almost independent of task dependent language resources and robust to term variations, making it highly portable to various information retrieval and natural language understanding tasks.

1. INTRODUCTION
In natural language understanding and information retrieval, it is well known that literally matching terms in documents with those of a query can be a problem. This is because there are usually many ways to express a given concept (synonym), and the literal terms in a query may not match those of a relevant document. Latent Semantics Indexing (LSI) is a powerful approach and successfully applied to information retrieval [1], natural language understanding and, particularly, to the problem of natural language call routing [2,3,4].

LSI is a vector space based approach. It assumes that there is some underlying or latent structure in the usage of terms in the document and query, which is obscured by the variability of the language and the choice of terms. It represents the terms (features) and categories as vectors in a semantic space. This representation is derived from a truncated singular value decomposition (SVD) on the term-category matrix constructed from the labeled training data. An unknown query is mapped into this semantic space as a vector, and its category can be determined based on the similarities between the query vector and its closest category vector.

The construction of term-category matrix is the first step of implementing LSI based classifier. In the term-category matrix, each term maps to a unique row vector and each category maps to a unique column vector. The matrix is often large and highly asymmetric with many times more rows (terms) than the number of columns (categories). The accuracy of a LSI classifier is therefore subject to the selection of terms. In addition to accuracy, term selection for large data set can significantly improves the runtime performance. In the previous LSI based approach [2,3], terms are selected based on their occurrence statistics (or frequencies) in the training data. Terms with occurrence less than a pre-set threshold are thrown away. Terms selected or discarded in this process may or may not be salient. It relies on other language processing resources (e.g. stop word list, ignore word list, etc.) and LSI to find out latent structure between various terms. Although this process has the advantage of making the LSI classifier construction almost automatic, it nevertheless is based on the statistics of term occurrences not the discriminative power nor the statistical significance of the term in classification.

In this paper, we propose an approach of LSI in natural language call routing that is based on a discriminative term selection for improved performance and robustness. The discriminative term selection is based on the criterion of Information Gain (IG) [5,6,7]. The discriminative power of the term is measured by the average entropy variations on the categories when the term is present or absent. The proposed approach was tested on several natural language call routing tasks. The experimental results indicate that significant performance gain in terms of precision and recall can be obtained from the proposed approach. The performance gain is most significant for small or medium size training corpus, where performance gain as much as 27% on precision and 26.5% on recall over the baseline LSI approach is observed. Moreover, the proposed approach can remove some critical language dependent resource in LSI based natural language call router, making it more robust and portable for different tasks and even for different languages.

2. Information Gain in Term Selection
Term Selection is an active research area in statistical text classification and machine learning [5]. Apart from approaches that require subjective knowledge, many term selection methods employ an iterative searching process between the terms and classifier. For large number of terms, the cost of term
Selection process can be very high and the selected terms are often tied to particular classifier.

More practical methods of term selection are often based on direct evaluation of the salient nature of the term. In this approach, each term is assigned a numeric value that indicates the importance of the term. A subset of terms can be chosen based on the value of this importance factor. Criteria in this class include Information Gain (IG), Mutual Information (MI), and \( \chi^2 \)-test [5,7]. Other techniques can also be used such as RELIEF and FOCUS [5]. Major advantages of this class of term selection methods are:

- The term selection algorithm can be made automatic or semi-automatic;
- The term selection process is usually fast as it involves one-pass computation of the training data.

Among these term selection criteria, IG based term selection is very unique. It measures the significance of the term based on the entropy variations of the categories, which relates to the perplexity of the classification task. In [7], a comparative study was performed and it indicates that IG is more effective for aggressive term removal without losing the classification accuracy.

The IG score of a term \( t_i \), \( IG(t_i) \), is calculated according to the following formulas:

\[
IG(t_i) = H(C) - H(C | t_i) - H(C | \overline{t}_i)
\]

(1)

\[
H(C) = - \sum_{j=1}^{n} p(c_j) \log(p(c_j))
\]

(2)

\[
H(C | t_i) = - p(t_i) \sum_{j=1}^{n} p(c_j | t_i) \log(p(c_j | t_i))
\]

(3)

\[
H(C | \overline{t}_i) = - p(\overline{t}_i) \sum_{j=1}^{n} p(c_j | \overline{t}_i) \log(p(c_j | \overline{t}_i))
\]

(4)

where \( n \) is the number of categories, and

- \( H(C) \): the entropy of the categories
- \( H(C | t_i) \): the conditional category entropy when \( t_i \) is present
- \( H(C | \overline{t}_i) \): the conditional entropy when \( t_i \) is absent
- \( p(c_j) \): the probability of category \( c_j \)
- \( p(c_j | t_i) \): the probability of category \( c_j \) given \( t_i \)
- \( p(c_j | \overline{t}_i) \): the probability of \( c_j \) without \( t_i \)

From the information-theoretic point of view, The IG score of a term is the degree of certainty gained about which category is “transmitted” when the term is “received” or not “received.” The right side of Formula (1) can be transformed to the following formula:

\[
\sum_{j=1}^{n} \left[ p(t_i, c_j) \log \left( \frac{p(t_i | c_j)}{p(c_j | t_i)} \right) + (p(c_j) - p(t_i, c_j)) \log \left( \frac{p(c_j) - p(t_i, c_j)}{p(c_j)(1 - p(t_i))} \right) \right]
\]

where we have:

- \( p(t_i) \): the probability of term \( t_i \)
- \( p(t_i, c_j) \): the joint probability of \( t_i \) and \( c_j \)

Multi-variate Bernoulli model described in [6] can be applied to estimate these probability parameters from the training data. In particular, \( p(c_j) \) is the number of document labeled to category \( c_j \) divided by the total number of documents; \( p(t_i) \) is the total number of documents containing term \( t_i \) divided by the total number of documents; and \( p(t_i, c_j) \) is the number of documents labeled \( c_j \) and containing term \( t_i \) divided by the total number of documents.

3. Information Gain Based LSI Algorithm

In this section, we describe the approach and implementation of an IG enhanced LSI classifier. The focus is on the IG extension part in the proposed approach and we refer to [2,3] for other details of LSI based classifier.

The training corpus for LSI based classifier is a collection of documents with corresponding categories labeled. It is first processed by a linguistic analysis module to convert words in the document into a sequence of raw terms. This module often based on morphological rules, such as Porter stemming, and linguistic resources such as root dictionary, ignore word list and stop word list, etc. The terms used in LSI analysis can be based on term unigram, bigram and trigram that correspond to raw terms, raw term pairs and raw term triplets. The inclusion of terms based on raw term sequence such as bigram and trigram, in LSI analysis is an effective way to improve the classifier performance and alleviate the limitation of the bag-of-word model used in LSI approach. However, this can lead to a huge increase of the dimension of the term-category matrix, because of the exponential growth due to the length of the raw term sequences. This makes the issue of term selection even more acute.

In the conventional approach, the term-category matrix \( M \) is formed by terms which occur more than a pre-set threshold in the training corpus. Each selected term is mapped to a unique row vector and each category is mapped to a unique column vector in the term-category matrix. The \( M[i,j] \) cell of the term-category matrix is the sample count that the \( i \)-th term occurs in \( j \)-th category. It is often advantageous to weight the raw counts to fine tune the contribution of each term in vector based content analysis. In our implementation, \( M[i,j] \), is assigned a weight based on \( TF*IDF \). An \( m \times k \) term matrix \( T \) and a \( n \times k \) category matrix \( C \) are derived by decomposing \( M \) through the SVD process, such that row \( T[i] \) is the term vector for the \( i \)-th
term, and row $C[i]$ is the category vector for the $i$-th category as typical in LSI based approach [2].

In our proposed approach, an additional module of IG based term selection is implemented. Instead of using terms based on their frequencies, terms are selected and used in the term-category matrix based on their discriminative power according to IG criterion. It consists of the following steps: (1) sort the terms by their IG values in a descending order; (2) select top $p$ percentile of terms according to the IG score distribution; (3) construct the term-category matrix and perform LSI analysis based on terms selected from (2). Since more than one term can have the same IG value, we take terms belong to the least upper bound of top $p$ percentile of IG score.

To categorize an unknown document, a query vector $X$ is derived from these terms as in LSI according to IG enhanced term-category matrix. Each category is assigned a similarity score against $X$ using $\cos(X, C[i])$. The n-best categories can then be chosen for $X$.

4. EXPERIMENTAL SETUP

The proposed approach is applied to the task of natural language call routing (NLCR). Call routing is a natural language understanding task of directing the user’s call to the appropriate destination within a call center, or providing auto-response or key information according to the user’s request. It has various applications in customer relation management. NLCR is similar to topic identification and document routing because it is to identify one of $n$ topics (destinations) most closely matches a user’s request. However it often distinguishes from these document applications by the requirement of single destination for routing, the variations in spoken inputs should it be a voice query, and the interactive nature that allows the query to be refined in a subsequent dialogue. In this paper, we focus on the classification part of an NLCR system and compare our approach with term frequency based conventional LSI approach.

In all experimental studies, recall and precision are the metrics used to measure the performance of the classifier. For each classifier, recall is defined to be (total number of correctly classified document)/(total number of documents), and precision is (total number of correctly classified documents)/(total number of classified documents).

The experiments were performed on a call routing task based on the transcriptions of spoken dialog data. The training session of the database consists of 3510 training documents in 23 categories. The independent test session consists of 307 test documents in 21 categories (2 categories are not observed). The average document length, counted by words, is 8.1 in the training set and 14.5 words in the testing set. The number of documents in each category is highly unbalanced. The standard deviation of number of documents in each category is 33.09 for the test set and it is 393.76 for the training set.

In order to study the behavior of the classifier for different size of the training data, two sub-training sets were created from original training data of 3510 training documents. One subset has 1755 training documents and the other one has 1404 documents. There were total of three training sets used in the experiments. Two separate experimental studies were performed. One is to study the performance advantages and the other one is to study the robustness of the proposed approach.

4.1. COMPARATIVE STUDY

Three sets of comparative experiments were performed. They correspond to the following three approaches:

- Baseline: all terms from language processing module are selected for LSI based classifier.
- Term Counts approach: select terms from language processing module if the occurrence counts (frequencies) are above some threshold.
- The proposed IG based approach: select terms from language processing module according to IG criterion.

Experimental results on relative performance improvements over the two comparative approaches are tabulated in Table 1 using different size of training data. The experimental results indicated that the proposed approach has the performance advantages in all three conditions over the other two approaches. It is important to point out that the dimension of the term-category matrix in each approach is very different. Baseline approach has the largest (4735) term dimension and IG based approach has the smallest (720) term dimension. The most significant performance gain was observed when training data were sparse. A performance gain of 27.4% in precision and 26.5% in recall were observed over the baseline on training set with 1404 samples. Similar performance gain was observed for the term counts based approach (22.5%, 22.4%).

<table>
<thead>
<tr>
<th></th>
<th>IG over Baseline</th>
<th>IG over Term Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>precision</td>
<td>recall</td>
</tr>
<tr>
<td>3510</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>1755</td>
<td>26.5</td>
<td>25.5</td>
</tr>
<tr>
<td>1404</td>
<td>27.4</td>
<td>26.5</td>
</tr>
<tr>
<td>avg</td>
<td>18.6</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Table 1: Relative error rate reduction of IG approach

4.2. PORTABILITY AND ROBUSTNESS STUDY

Experiments were performed to study the robustness and portability of the proposed IG based LSI approach. The LSI classifiers constructed in this study were based on the same three approaches described in the previous subsection. But it distinguishes from the previous study in that all LSI classifiers were built directly from the raw terms without task dependent language resources. Except a generic Porter stemmer, all stop list, ignore list, root list were not used. In the baseline case, the total number of term dimension increased from 4735 to 38724. Without the language dependent resources, terms are noisy, presenting a challenge to the classifier design.

Results of this study for three different size training data sets are tabulated in Table 3. It is interesting to note that the performance gain of the proposed approach is even more significant in the condition of no other language resources. The
performance gain is cross board on all conditions. Comparing with the baseline, the relative performance gains of 43% to 61% were observed. Comparing with the term count based approach with about the same number of terms, the performance gain was even more significant and relative improvements in the range of 60% to 71% were observed. Most importantly, the performance of the proposed approach was almost not affected by those language resources dependent filtering. It indicates that the proposed IG enhanced LSI approach is not only robust but also portable to different applications.

Table 2: LSI without root, ignore and stopword lists

<table>
<thead>
<tr>
<th>Train</th>
<th>Metric</th>
<th>Baseline</th>
<th>Term Count</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>3510</td>
<td>selection</td>
<td>all</td>
<td>≥20,20,20</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>terms</td>
<td>38724</td>
<td>789</td>
<td>775</td>
</tr>
<tr>
<td></td>
<td>precision</td>
<td>83.38</td>
<td>77.19</td>
<td>93.48</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>83.38</td>
<td>77.19</td>
<td>93.48</td>
</tr>
<tr>
<td>1755</td>
<td>selection</td>
<td>all</td>
<td>≥15,16,16</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>terms</td>
<td>22001</td>
<td>498</td>
<td>441</td>
</tr>
<tr>
<td></td>
<td>precision</td>
<td>80.45</td>
<td>68.4</td>
<td>88.92</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>80.45</td>
<td>68.4</td>
<td>88.92</td>
</tr>
<tr>
<td>1404</td>
<td>selection</td>
<td>all</td>
<td>≥8,9,9</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>terms</td>
<td>18271</td>
<td>732</td>
<td>733</td>
</tr>
<tr>
<td></td>
<td>precision</td>
<td>78.5</td>
<td>72.63</td>
<td>88.92</td>
</tr>
<tr>
<td></td>
<td>recall</td>
<td>78.5</td>
<td>72.63</td>
<td>88.92</td>
</tr>
</tbody>
</table>

Further study on the sensitivity of IG threshold to the classifier performance was performed. Figure 1 is a plot of the accuracy change with different IG threshold for LSI classifier trained on the full 3510 training samples. 100% point corresponds to the baseline case where all terms are used in LSI and 1% point corresponds to the LSI classifier based on the terms with top 1% IG score. Because more than one term can have the same IG score, the percentile of the size of the terms is always greater or equals to the IG percentile.

5. SUMMARY

We presented in this paper an approach that combines an information gain based discriminative term selection process in LSI based classifier design. In IG based term selection, the discriminative power of the term is measured by the entropy variation averaged over all categories, conditioned upon whether the term is present or absent. The LSI classifier is based on the salient terms according to IG criterion. Compared to the salience measure proposed in [8,9], IG is more general as it makes use of term absence information. Because of this, the discriminative power of low frequency terms can still be measured despite their absence. The accuracy of an IG based LSI classifier on small amount of training samples indicates empirically that IG based approach is robust with low frequency terms. Various experimental studies were performed. Comparing to the conventional LSI approach based on term frequencies, significant performance improvements were observed. Most importantly, the proposed approach is robust and almost independent of language specific resources, making it portable to different tasks. Our results are comparable to the best LSI results on call routing using this database, and IG computation is quite fast with a computational complexity which is linear to the product of terms and categories.

6. REFERENCES