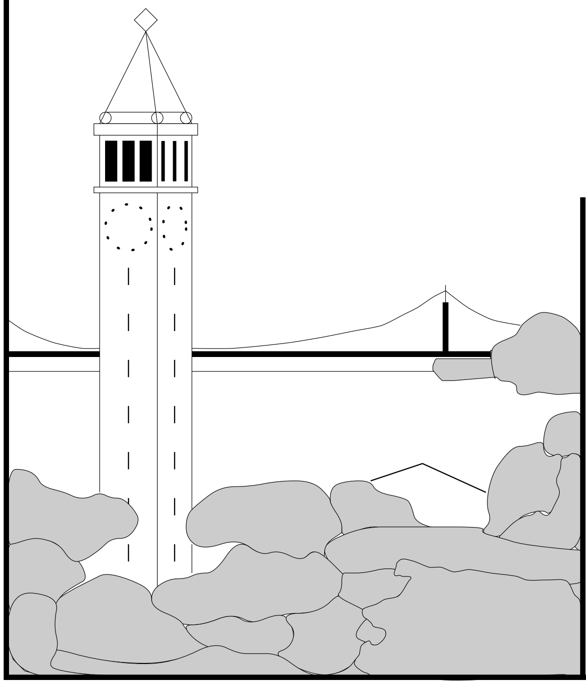
Topic Characterization of Full Length Texts Using Direct and Indirect Term Evidence

David E. Fisher



Report No. UCB/CSD 94–809

May 1994

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Abstract

This project evaluates two families of algorithms that can be used to automatically classify general texts within a set of conceptual categories. The first family uses *indirect evidence* in the form of term–category co-occurrence data. The second uses *direct evidence* based on the senses of the terms, where a term's senses are designated by the categories that it is a member of in a thesaurus. The direct evidence algorithms incorporate varying degrees of indirect evidence as well.

For these experiments a set of 3864 conceptual categories were derived from the noun hierarchy of WordNet, an on-line thesaurus. The co-occurrence data for the associational and disambiguation algorithms was collected from a corpus of 3,711 AP newswire articles, comprising approximately 1.7 million words of text. Each of the algorithms was applied to all of the articles in the AP corpus, with their performance evaluated both qualitatively and quantitatively.

The results of these experiments show that both classes of algorithms have potential as fully automatic text classifiers. The direct methods produce qualitatively better classifications than the indirect ones when applied to AP newswire texts. The direct methods also achieve both a higher precision, 86.75% correctly classified (best case) versus 72.34%, and a higher approximate recall.

The experiments identify limiting factors on the performance of the algorithms. The primary limitations stem from the quality of the thesaural categories, which were derived automatically, and from the performance of the term sense disambiguation algorithm. The former can be addressed with human intervention, the latter with a larger training set for the statistical database.

Contents

1 Introduction

To facilitate information retrieval one would like to be able to classify documents based on their content, rather than just by the terms they contain. One such classification system is the Library of Congress Subject Headings (LOCSH); another is one based on the category classifications of words from a thesaurus. These have been used in automatic document classifiers (labelers), by Larson [7], processing titles and subject headings, and by Liddy [8], Hearst and Schütze [5], and Hearst [6], processing full texts. The full text algorithms use two different types of evidence for selecting the labels to assign to a text, *direct* and *indirect*. Direct evidence uses a mapping from terms onto a category set, labeling a text with some combination of these categories. Indirect evidence uses associations between terms and categories, in the form of co-occurrence data, labeling a text with some combination of the categories that co-occur with the terms of the text.

This research examines the performance of the associational (indirect evidence) approach compared to a number of variations on the use of direct evidence, and also to an algorithm that combines both direct and indirect evidence. The algorithms use a set of 3864 conceptual categories derived from WordNet [9], an on-line thesaurus, using Hearst's [5, 6] algorithm. The associational algorithm is based on Yarowsky's [10] disambiguation algorithm, as it was employed by Hearst [5, 6]. Combining these components provides a mechanism for performing completely automatic text classification.

The approaches are motivated by the intuition that the content of a text can be approximated by some set of the categories of the terms that make up that text. There are numerous ways in which the meaning of an utterance exceeds this ideal: the relation between elements, inferences, metaphors, and idioms are some examples. However, the goal of these algorithms is not to completely understand a text, but rather to position it within the space defined by a conceptual hierarchy.

Unlike single label classifiers, these algorithms permit assigning multiple categories to a text. The category assignments situate the text within the conceptual hierarchy, allowing it to be retrieved directly. The assignments can also be used for "browsing" a collection of texts. In this case texts that are assigned categories that are near each other in the hierarchy will be close together in the browser.

The categories derived from WordNet provide conceptual labels for the model. The disambiguation algorithm provides a method for choosing the senses to be combined. The training set and test bed come from a collection of 3711 Associated Press newswire articles that are not restricted with respect to domain.

In the light of both quantitative and qualitative analysis, these algorithms demonstrate good potential for completely automatic classification of texts. The performance of the direct evidence methods is superior to that of the indirect evidence. Overall the performance is not as good as it could have been, due to training with too little text. The results do justify further experiments using a larger training set.

The remainder of this paper is laid out as follows. I begin by describing the construction of the category set from WordNet, followed by a description of the AP corpus. This is followed by descriptions of the associational training and the algorithms being considered. The performance of both the support components (category set, disambiguation) and the labeling algorithms is then presented. I conclude with some directions for future exploration.

2 Thesaural Categories

This research uses a set of thesaural categories constructed from WordNet (v1.4) [9], a large on-line thesaurus. WordNet classifies words by membership in *synsets*, which are collections of synonymous terms. These synsets are broken down by part of speech; this research uses only the nouns. In addition to the terms themselves, a synset contains a list of the relations that it participates in, such as, in the case of the nouns, hyponymy — hypernymy.

This is the relation used by Hearst and Schütze's algorithm for deriving conceptual categories. Their algorithm traverses the WordNet noun hierarchy creating categories from synsets according to the following constraints. If the number of descendants of a synset (terms and subordinate terms) is greater than a lower bound and less than an upper bound, that synset and its descendants are assigned to a category. When the lower bound is not exceeded, the algorithm moves up the hierarchy. When the upper bound is exceeded, the algorithm splits off the descendants. In the case of a leaf node that exceeds the upper bound, i.e., there are no children to split, the node is made a category.

For these experiments I used upper bound on category size of 15 and a lower bound of 5. These parameters produce a set of 3864 categories, averaging 18.34 terms in each category. The choice of category size was motivated by the desire to make fine grained classifications, leaving open the possibility of performing additional processing to navigate the conceptual hierarchy at a later time. Hearst [5, 6] chose the other alternative, that of abstracting to a much greater degree, using fewer categories (726 and 106). The degree of distinction for these categories varies, because the splitting criterion is based solely on size. I explore the merits/demerits of the category set in Section 7.1.

3 AP Newswire Corpus

This training set and test bed consist of 3,711 Associated Press newswire articles from the Tipster AP corpus[1]. The articles are a general slice of the AP newswire, dating from Jan. 1, 1989 through Jan. 17, 1989. The corpus contains approximately 1.7 million words, averaging 458 words per text. It has 47,287 tokens, that is, distinct words, including morphological variants.

Because the thesaural categories were constructed from the WordNet noun hierarchy, the algorithms that employ direct evidence needed to consider those terms that were ambiguous with respect to part of speech only when they were acting as nouns. To do this I used PARTS [1], a stochastic part of speech tagger, to tag each of the texts in the corpus.

4 Association Training

Both the disambiguation and the associational labeling algorithm require term-to-category co-occurrence frequencies. These frequencies are collected in the training phase, which involves two passes over the training corpus. The first pass counts the terms in the corpus; the second counts the co-occurrences. A stop-list of 908 words is used to filter out function words, and other content-less terms.

[1]The author would like to thank Donna Harman who made this collection available to the Berkeley Full-Text Retrieval Research Group, a participant in the DARPA-sponsored TREC conference.

Let $f(term)$ be the number of times *term* occurs in the training corpus, M be a frequency threshold, W be a fixed length window surrounding a target term, *target*, and let $\alpha(C,t)$ be the association metric between a term and a category. $\alpha(C,t)$ is computed for each term in the training corpus by doing the following.

The window W slides over each training text, updating the co-occurrence frequencies each time the term in the target position is not a stop word. Terms that do not have a definition in the lexicon are ignored. The formula for updating is:

foreach $cat \in \text{senses}(target)$ **do**
    foreach $term \in W$ **do**
        if $f(term) \geq M$
            $\alpha(cat,term) \leftarrow \frac{1}{f(target)} + \alpha(cat,term)$
        fi
    od
od

Note that the association metric is normalized by the frequency of the term in the training corpus. This is the strategy used by both Yarowsky and Hearst to prevent frequent terms from dominating infrequent ones. For these experiments I use a 101 term window, 50 terms on either side of the target. The frequency threshold, M, is supposed to filter low frequency terms out of the statistics and is used as an alternative to the smoothing algorithm employed by Yarowsky. This research used $M = 4$.

Training was performed on a DEC Alpha AXP workstation, configured with 64 Megabytes of memory. It took approximately 65 hours of real time to complete the training, utilizing 75 CPU minutes. The table of $\alpha(C,t)$ values required 385 Megabytes of virtual memory address space and 250 Megabytes of disk space for storage. Most of the real time needed for the training was spent waiting for NFS disk accesses. None of the code used for training was optimized, either with respect to execution speed or storage requirements.

5 Term Sense Disambiguation

The disambiguation algorithm is based on that of Yarowsky [10]. It uses the $\alpha(C,t)$ from the training phase, described above, to determine which sense of the target term is being used. Yarowsky defines an estimate for the *salience* of a term for a given category as $\frac{P(term|cat)}{P(term)}$, the probability of the term appearing in the context of the category divided by the probability of the term in the whole corpus. This measure, in a similar fashion to the mutual information statistic, approximates how good an indicator a term is for a category. The log of the salience estimate multiplied by the probability of the category ($P(C)$) provides the evidence term for the disambiguation. As was done by Yarowsky [10] and Hearst and Schütze [5], the categories are assumed to be uniformly distributed, so $P(C)$ is omitted from the computation.

The algorithm disambiguates a target term as follows.

foreach $cat \in \text{senses}(target)$ **do**
    foreach $term \in W$ **do**
        if $f(term) \geq M$
            if $\log\left(\frac{P(term|cat)}{P(term)}\right) > 0$ **then**
                $\text{Votes}[\text{index}(cat)] \leftarrow \log\left(\frac{P(term|cat)}{P(term)}\right) + \text{Votes}[\text{index}(cat)]$
            fi
        fi
    od
od
$\text{Sense}(target) \leftarrow cat : \text{Votes}[\text{index}(cat)] = \max_C \text{Votes}[C]$

The evidence for each of the categories is collected for each of the terms in the window surrounding the target. When the log of the salience is less than zero, resulting from the salience of the term for the category being less than one, the negative evidence is not used. This is because negative evidence is more likely to be noisy than positive evidence. A term does not necessarily provide evidence against a category just because it does not provide evidence for it.

This algorithm was applied to the AP corpus, with its output used to construct term sense prior probabilities database used in two of the direct evidence algorithms. Running on the same machine as the training, it took approximately 4900 CPU minutes (90 hours real time) to disambiguate the 1.7 million word corpus. Unlike the training, which was I/O bound, disambiguation was compute bound. As was the case for the training, the implementation has not been optimized with respect to execution speed or memory requirements.

6 Topic Labeling

Each of the topic labeling algorithms presented here is a variation on the disambiguation algorithm. What distinguishes them is where they get their evidence from. I first present the associational approach, followed by each of the direct evidence methods.

6.1 Indirect Evidence

The disambiguation algorithm adapts readily to one for topic labeling. The algorithm used here is based directly on that of Hearst and Schütze [5], using a different formula for combining the associational information from the terms in a text. In that research they updated the category vector once every 30 terms, so each term in the 100 word window would contribute three times. This method of probing is like disambiguating every thirtieth term, and as such does not seem to be making the best use of the evidence available.

In this research I take the simple alternative of using all of the evidence that is available. The computation proceeds as follows: For each term in the text, the degree of association with each of the categories is computed, just as in disambiguation. The vectors for each of the terms are summed to produce the category vector for the text. Each term contributes once to the classification of the text. The only difference from the disambiguation computation is that, rather than constraining the candidate set of categories to the senses of the the term in the target position, every category is considered.

While using all of the evidence is intuitively more appealing than only using some, this simple combination does not make any attempt to filter the spurious categories from the actual category being used in the target position at the time of each update. How to do this filtering without having the computation degrade to the disambiguation algorithm is an open question. That is to say, the classification of the text should not be constrained to only include the categories that appear directly in the text, but rather, this algorithm should capture those categories that are not used specifically.

6.2 Direct Evidence

There are three variations of the direct evidence labeling algorithm. The first is the base algorithm, closely related to that of Liddy and Paik [8].

```
foreach term do
    if content_term(term) then
        foreach cat ∈ senses(term)
            Votes[index(cat)] ← Votes[index(cat)] + evidence(term, cat)
        od
    fi
od
```

Figure 1: Direct Evidence Labeling Algorithm

The other two are refinements of the first, each using a weighting strategy intended to produce more accurate labelings. In the direct approach, each content term in the text contributes evidence for each of the categories that it has as one of its senses. Content terms are those nouns that are not on the stop-list and are in the WordNet derived lexicon. The evidence from all of the terms is summed, and the resultant vector of categories is sorted. The top ten ranked categories are assigned to the text. The choice of taking the top ten, as opposed to five or fifteen, was arbitrary. All of the top ten are examined in the qualitative analysis; only the top three are considered for the quantitative analysis.

6.2.1 Base Algorithm

Figure 1 shows pseudocode for all of the direct evidence labeling methods. In its simplest form

$$\forall term\, \forall cat \in \mathtt{senses}(term) : \mathtt{evidence}(term,\ cat) = 1$$

In this case, each polysemous term is treated as one occurrence of each of its senses. While this provides a baseline for the categories used in a text, it can not be correct, as only one of the senses of each term was intended by the author of the text. This observation motivates weighting the evidence. There are two ways that the evidence from polysemous terms can be weighted, applying the uniform distribution assumption, and using the prior probabilities of the term's senses. Each of these strategies are described below.

6.2.2 Weighting with the Uniform Distribution Assumption

No additional information is required to apply the uniform distribution assumption. Then, $\texttt{evidence}(term,\ cat) = \frac{1}{\texttt{numsenses}(term)}$. This approach dilutes the contribution of polysemous terms, but still allows unintended senses of a term to contribute to the topic labeling. It provides a poor model for those terms that have an extremely common primary sense, and some number of rarer secondary senses. However, with these caveats in mind, it does provide a reasonable model of the categories appearing in a text in the absence of additional information.

6.2.3 Weighting with Prior Probabilities

A better model of the categories present in a text is one that weights the evidence for a term's senses by their prior probabilities. This is consistent with Gale et al's [3] observation that disambiguating a term by always assigning it its most frequent sense achieves 92% correct assignments. So, in this approach, $\texttt{evidence}(term,\ cat) = P(\texttt{sense}(term) = cat)$. The problem, then, is where to get the priors.

One way to collect the priors is to count the sense usages for each term in a corpus. Unfortunately, the terms in the AP corpus are not tagged with respect to which of their senses is being used. So, to approximate this data, I do the following:

1. Run the disambiguation algorithm on each term of the corpus, recording which sense it selects.
2. Take these frequencies for each term as the priors, normalizing by the number of occurrences of the term (that were disambiguated).

This frequency data contains noise from two sources. First, not every occurrence of each term is disambiguated. Second, not every disambiguation decision is correct. I discuss the impact of these errors in Section 7.2. Even with the noise, this is more information than is available with the uniform distribution assumption, and it should produce a better model of the content of a text.

6.3 Combined Evidence

The final approach attempts to create an even better model of a text by combining the indirect and direct evidence. Here if a term can be disambiguated, that category gets a vote of 1 and all of the other senses of the term receive no vote. If a term is not disambiguated, then use the evidence from the prior probabilities. Intuitively this is the most appealing model, assuming disambiguation selects the intended senses of the terms, for the content of a text. With this approach, unintended senses do not contribute to the topicalization of the text, except in those cases where disambiguation fails.

Each of the algorithms described was applied to each article in the AP corpus. The running times for all of them averaged less than one minute per text of real time, with the associational algorithm typically taking the longest. This resulted in mountains of data which I now endeavor to analyze.

7 Results and Evaluation

Before evaluating the performance of the labeling algorithms, I consider the support components and their effect on the results reported here. I begin with the category set derived from WordNet, which does suffer from some limitations. I then evaluate the performance of the disambiguation algorithm, which is critical for both the associational labeling and the direct methods that use either the term sense priors, or the disambiguator output. Then I present both a qualitative and a quantitative evaluation of the labeling algorithms.

7.1 Thesaural Categories — Grades of Distinction

One of the problems with automatically constructing a set of categories from WordNet is the uneven granularity that results. The algorithm's primary goal is to collect terms into sets using the size of the set as the criterion for splitting/joining categories. This results in some extremely fine-grained categories, such as one for each of a number of varieties of mushrooms, and some coarser categories, such as SOCIAL-SCIENCE, which spans criminology, demography, economic, political science, econometrics, sociology, and geopolitics. In some cases the fine distinctions capture topical differences that are useful in textual classification. In others, such as in the case of the term "film," the distinction distracts from the meaning of the term.

Film has five senses in the lexicon:

0.640	00865	movie film picture (MOVIE)
0.151	01068	media mass_media (MASS-MEDIA)
0.058	01995	wrapping wrap wrapper (SARAN-WRAP)
0.138	02245	photographic_material (PHOTO-FILM)
0.013	03713	object inanimate_object (POND-SCUM)

The first two of these, MOVIE and MASS-MEDIA, cover individual motion pictures and motion pictures as an art form, for example, *the film "Platoon,"* and *Film is a very powerful art medium.* From the point of view of classification it is less important to make this distinction, as individual movies are instances of the art form, than it is to make the distinction between the those senses and the SARAN-WRAP or the POND-SCUM senses.

A second problem is that WordNet uses separate hierarchies for the different parts of speech. This research uses only noun information, which is intuitively less informative than the information that could be gleaned from all of the words in a text. One alternative would be to manually merge the different parts of speech from WordNet, a daunting task to be sure. Another would be to appeal to a "better" thesaurus. Unfortunately, no such better thesaurus is currently available in a machine readable format. In each of the cases, the quality of the categories could be improved by manual intervention.

7.2 Term Sense Disambiguation

I evaluated the performance of the disambiguation algorithm by selecting three texts from the AP corpus, applying the algorithm, and manually classifying the results. Only those terms that were actually used as nouns in the texts were evaluated; the ones which were misclassified by PARTS, primarily noun-verb ambiguities, were discarded. Terms with a single sense are regarded as always correct. For the polysemous terms, a sense is labeled incorrect in those cases where one of its alternative senses is more appropriate for the sentence in which it occurs. This criterion works well for disparate senses, such as the MOVIE versus the SARAN-WRAP senses of the word "film." When the senses are very similar, such as MOVIE versus MASS-

Baseline

	Single Sense Terms	%	Polysemous Terms	%	Total Terms	%
Correct	168	100.00	128	46.55	296	66.82
Incorrect	0		147		147	
Total	168		275		443	

Algorithm

	Single Sense Terms	%	Polysemous Terms	%	Total Terms	%
Correct	168	100.00	193	70.18	361	81.49
Incorrect	0		82		82	
Total	168		275		443	

Figure 2: Disambiguation Performance

MEDIA for "film," there is a greater chance for human performance error. In every case I endeavored to give the algorithm the benefit of the doubt.

Gale et al [3] define a baseline algorithm for word-sense disambiguation that always classifies a term as its most frequent sense. This provides a lower bound on the performance that should be achieved by any alternative disambiguation algorithm. Figure 2 shows the results for both the baseline and the disambiguation algorithms.

On the face of it, this performance seems very poor. Gale et al [3] report 92% for all terms and 75% for polysemous terms, however there are a number of differences between their evaluation and mine. They randomly selected 97 words, 67 of which were unambiguous, and measured the performance of the baseline using the frequencies of occurrence in their hand-labeled test set to determine the most frequent sense of each term. They do not report the total number of terms used to compute their percentages. Additionally, term senses in their study were derived from 1042 Roget's Thesaurus categories as opposed to the 3864 categories for this study. That is to say, their baseline is making fewer discriminations.

Looking at only the 30 polysemous terms, they report a total of 84 senses. Those same terms have 107 senses in the lexicon constructed from WordNet.

Where they have 2.8 senses/term, I have 3.57. Clearly, with more categories to choose from, making a correct choice is harder.

For the 443 terms in the sampled texts there are a total of 1267 senses, averaging 2.86 senses/term. Of those 443, 275 are polysemous with 1099 senses for an average of 4.0 senses/term. So, if I randomly choose a sense for each term I can expect to do no worse than 34.97% for all terms, 25.0% for the polysemous ones.

Their baseline had the benefit of a hand-tagged training set, whereas my prior probabilities are based on the output of the algorithm being evaluated here. As seen in the totals, just over 70% of the polysemous terms are correctly disambiguated. The noise introduced by the errors is visible in its effect on the baseline performance, where just over 46% of the polysemous terms are correctly assigned. Be that as it may, the performance is still unsatisfying, especially compared to Yarowsky's [10] average of 92%.

One of the problems with this algorithm is that it requires a large training corpus from which to collect the association frequencies. Although I have shown in [2] that a corpus as small as 500,000 words can be used for a similar frequency based technique, that was in the context of a limited domain. The AP corpus used here totals 1.7 million words, which is small compared to the 10 million words used by Yarowsky [10] and the 8.7 million words used by Hearst and Schütze [5].

The second problem it faces is the quality of the thesaurus. Recall that the categories were constructed from WordNet using size as the selection criterion, producing differing levels of granularity. In the case of "film" performance suffers because two of its senses, MOVIE and MASS-MEDIA, split the vote, when actually those two senses should be merged into a single category. These problems become an issue for both the associational labeling, which is a similar algorithm, and the direct methods that use the output of the disambiguator, either directly or in the form of the prior probabilities on senses.

7.3 Topic Labeling

It is difficult to measure the performance of the various labeling algorithms quantitatively without hand classifying a test set of AP articles. The method I have chosen is to measure the precision (number of correct assignments out of those assigned) for a sample from the category set. This still requires a human judge, but the decision is less prone to error than that of choosing categories that apply from the full set of 3864. An approximation of the recall (number of correct assignments out of all that should have been assigned) can be obtained by combining the sets of correctly classified texts across the algorithms. This method will still miss those relevant texts that were not assigned a highly salient category by any of the algorithms, so the approximated recall is higher than the true recall. This is not a problem, as the ranking of the algorithms does not change if the number of relevant texts is increased.

Although there is an ordering in the output for each of the algorithms, it is not clear how to compare the ranks, either between different texts for a single algorithm or between algorithms on the same text. For the purpose of evaluating precision, the top three categories assigned to a text are taken as the classification, without looking at their ranks.

Sections A.1, A.2 and A.3 contain three sample texts from the AP corpus. The output of the labeling algorithms will be examined in a more qualitative fashion for those texts.

7.3.1 Qualitative Analysis

Figures 3, 4, 5, 6, and 7 show the top ten category assignments for three AP articles by each of the five algorithms. Each of the articles is given a brief gloss below.

AP890101-0001 Section A.1. This article discusses the spate of Vietnam Era (60's) movies from the late 1980's. A reasonable set of topic terms for this article would include; movies, Vietnam, war, the 60's, and civil rights.

AP890109-0002 Section A.2. This article describes a request for Canadian asylum by a Russian emigree accused of being a Nazi propagandist. Topic terms for this article would include; immigration hearing, asylum, deportation, and Nazi propagandist.

AP890112-0001 Section A.3. This article describes the outcome of an extradition request for a Salvadoran accused of assassinating an Archbishop. For this text, topic terms would include; extradition hearing, deportation, assassination, and El Salvador.

First, I discuss each algorithm's performance for the three texts, followed by a comparison across the algorithms.

Associational Evidence Figure 3 shows the output from the associational algorithm. The first text (AP890101-0001) is an example of a difficult text for the associational algorithm. The top category SOCIAL-SCIENCE [3208] includes "politics" and as such captures one aspect of the text, but I consider it to be more of a peripheral topic. The second and ninth categories, racket [2298] and sports_implement [2299], are two closely related senses of the term "bat" from the movie title "BAT 21" in the text. These categories are activated because there is no disambiguation in the associational training. Additionally, "bat" is infrequent in the training corpus, just reaching the minimum number of occurrences (4), with each of these an instance of the movie title. None of the senses in the lexicon actually apply to the usage in the title. This behavior demonstrates one of the problems that can occur when spurious associations are trained in. The third category, script [347], is reasonable in the same fashion as the first. The best choice from the top ten, however, is the last category, movie [865].

The second text (AP890109-0002) is another difficult text for similar reasons. The top two categories capture unintended senses for two terms in the text. The first, achromatic_color [178], comes from "Grey," the name of the attorney. This is an example of a problem with proper names. Unlike Supreme Court or Canada, Grey, as a name, should not contribute to the topic classification. Unfortunately, there is no way to distinguish meaningful proper names with the information in WordNet, so, rather than excluding content bearing proper names, spurious proper names are included. An alternative, that would require human intervention, is to augment the lexicon with content bearing proper names, ignoring all others. The second category, nervous_tissue [1767], results directly from a spurious sense of the term "tract," another low frequency term. As in the first text, the appropriate categories are far down in the list, banishment [845] (deportation) seventh, and writing [350] ninth.

Performance on the final text (AP890112-0001) is markedly different. Here the top three categories all apply (although the murder should probably be below the second and third), and only three of the top ten are inappropriate. The first, chisel [2307] resulted from the term "drove" being tagged as

Score	Category Number	Category Name
	AP890101-0001	
276.7062	03208	social_science
276.1323	02298	racket racquet
269.6271	00347	script book
265.1193	03582	graduate_school
257.0557	02822	commissioned_military_officer
254.8447	03695	molecule
250.6811	01932	army_unit
250.2760	02728	historian historiographer
247.6937	02299	sports_implement
247.3209	00865	movie film picture
	AP890109-0002	
157.2674	00178	achromatic_color
156.5529	01767	nervous_tissue
141.1330	00309	literary_composition literature
136.2723	03174	system
113.0014	00949	teaching instruction pedagogy
107.5613	02485	literary_study
105.7594	00845	banishment proscription
103.0946	00250	Slavic Slavic_language
98.8111	00350	writing writings
98.3614	01832	applicant candidate
	AP890112-0001	
102.5115	00773	murder homicide slaying
101.7981	02509	court tribunal
97.0290	00845	banishment proscription
96.3072	02307	chisel
96.2109	02821	general_officer
95.4540	00338	writ judicial_writ
90.6745	03741	exemption immunity
90.2818	00749	traffic
89.6724	00397	evidence
89.2474	00889	inquiry enquiry

Figure 3: Labeling with Associational Evidence

a noun, rather than a verb, in both training and testing. The second, general_officer [2821], is a second proper name problem. In this case the name is "Napoleon." The last, traffic [749], is not completely inappropriate, as it is related to the arrest that precipitated the deportation hearing. Overall, this text is a good example of the kind of classification we would like to achieve by using the associational algorithm.

All three of the texts demonstrate that low frequency terms/categories can dominate the behavior of this algorithm, often producing poor results. A larger training set would help to offset the problem, as would disambiguation in the training phase.

Unweighted Category Counting The output is shown in Figure 4. This approach identifies a number of the problems that result from not disambiguating the terms in a text, and also from not weighting the contributions for polysemous terms. In the first text the top eight categories are the eight senses of the term "time." Clearly all eight should not get a vote each time "time" (or the other time terms) appears in the text. The time categories also illustrate one of the deficiencies of the thesaurus. Although there is evidence for the senses of the time terms, the text is not about time. Rather the time terms are providing a context for the content of the text. This is similar to the use-mention problem, where mentioning a term, such as "the word murder, from the Latin ...," does not mean the same thing as using the term, as in "the murder of the nuns ..." One meaningful category does make it into the top ten, movie [865], as the words movie and film are frequent in the text.

In the second text, the top four categories are related to the content of the text, but they do not provide a very satisfying characterization. In first, organic_phenomenon [3840], is one of the senses of death, but it is not the type of death the subject of the article faces if he is deported. Writing [350] comes from the propaganda, tracts, and articles. Status [3861] results from refugee status, which also produces migrant [2782], the only sense of "refugee." So, with some interpretation, this could be called a better characterization than that of the first text, but it does not provide any really useful information.

As was the case for the associational algorithm, performance is best on the final text. The second and fourth ranked categories are two senses of "request." Only one type of request is made in the text, so only one of these

Score	Category Number	Category Name
		AP890101-0001
18	03738	happening occurrence natural_event
14	00635	time_period period period_of_time
10	00039	time
10	00221	measure measurement
10	00609	time age
10	03336	moment instant
10	03357	datum data_point
10	00184	sound_property
8	03805	people
7	00865	movie film picture
		AP890109-0002
4	03840	organic_phenomenon
4	00350	writing writings
4	03861	status social_state
3	02782	migrant
3	03839	motivation motive need
3	03337	end ending finale finis finish
3	00635	time_period period period_of_time
3	01062	press public_press
3	00464	status social_rank social_class
3	00290	language linguistic_communication
		AP890112-0001
6	02587	government authorities regime
5	01107	request asking
4	00773	murder homicide slaying
3	00353	request petition solicitation
3	02487	dominance ascendance ascendence
3	00394	assertion averment asseveration
3	02509	court tribunal
3	00373	evidence
3	01201	lawyer attorney
3	01146	administration governance government

Figure 4: Unweighted Direct Evidence

categories belong in the classification. It does, however, take less interpretation to glean that this text is about some type of legal proceeding involving a murder.

Sometimes using a simple algorithm results in simple results, and sometimes it results in simple-minded results. This algorithm is one of the latter. By permitting multiple senses of a term to weigh in as heavily as the single sense of an unambiguous term, the algorithm allows spurious senses to rise to the top of the rankings. It does provide a baseline that any weighted strategy would need to exceed in order to be of any value.

Uniform Distribution Assumption Weighting The output is shown in Figure 5. In the absence of prior probabilities for the distribution of term senses, uniform weighting of the senses gives a form of disambiguation. With this method there is a marked improvement over the unweighted approach, as is seen in the first text. Asian_country [1155] has risen to the top. This is the only sense of the term "Vietnam." The second through fourth are questionable, with the categories people [3805] and happening [3738] in a similar class as the time terms. These are more content bearing than "time," but still bring little discrimination to the classification. It is difficult to decide if the movie title "Platoon" should be considered as evidence for army_unit [1932]. The movie that it names is about that topic; however, the movie should dominate. Below these come movie [865] and war [1005] which are definitely on point, but a little too low in the rankings. At ninth, right [642] comes from civil rights, also on point but too low.

The second text does not fare so well, demonstrating the proper name problem from Grey. The top three do indicate that the text is about a refugee associated with a university, but it is a stretch to read that into the categories. This text is the shortest of the three, and as such has fewer opportunities for an intersection between the categories for different terms in the text.

In the third text it is clear from the categories that someone is involved in a legal proceeding involving a government in North America. The murder [773] is in the top ten, but it is down at eighth.

Score	Category Number	Category Name
	AP890101-0001	
9.000000	01155	Asian_country Asian_nation
6.142857	03805	people
6.000000	01932	army_unit
5.916667	03738	happening occurrence natural_event
5.599999	00865	movie film picture
3.500000	01005	war warfare
3.333333	01875	city metropolis urban_center
3.000000	02586	writer author
3.000000	00642	right
3.000000	02758	citizen
	AP890109-0002	
3.000000	02389	university
3.000000	02782	migrant
3.000000	02758	citizen
3.000000	00178	achromatic_color
3.000000	01158	country state land nation
2.000000	01201	lawyer attorney
2.000000	01157	North_American_country
2.000000	01171	council
2.000000	00620	calendar_month month
2.000000	01189	conservative
	AP890112-0001	
4.333333	02509	court tribunal
4.000000	02758	citizen
3.083333	02587	government authorities regime
3.000000	01201	lawyer attorney
3.000000	01157	North_American_country
3.000000	01875	city metropolis urban_center
2.750000	00620	calendar_month month
2.500000	00773	murder homicide slaying
2.000000	02717	expert
2.000000	02821	general_officer

Figure 5: Uniform Distribution Weighted Direct Evidence

In each of the texts the performance is an improvement over the unweighted case. This algorithm provides a realistic benchmark. It requires no additional knowledge or training over the lexicon. The remaining methods, which bring additional information to bear, need to do better than this to be considered worthwhile.

Prior Probability Weighting Figure 6 shows the output. This strategy gets closer to capturing the intended senses of the terms in a text. Turning to the first text, the top four categories (ignoring the platoon problem) capture a great deal of the content of the text, movies about the Vietnam war. With the possible exception of army_unit [1932] none of the categories are inappropriate. Unfortunately, civil rights are still too far down the list.

In the second text the proper name problem recurs, and the overall characterization is not very different from that of the uniform distribution assumption method. The month [620] category is another of time type categories. Here the uses of "July" and "Jan." provide the time context for the content of the article, and should be interpreted as such.

In the third article the assassination moves up into third place. The presence of lawyer [1201] in the fifth position raises another question about interpretation. Although the article contains lawyers, who say a number of things, it isn't really about the lawyers; they are just players in the scene. This phenomenon also occurred when using the uniform distribution assumption weighting.

Overall the classifications are better with prior probability weighting than with uniform weighting, but not glaringly so. And, of course, the priors are somewhat suspect because they were derived from the output of the disambiguation algorithm, whose performance was less than stellar. Improvements to the disambiguation would improve the priors which should improve the performance of the classifications. But this algorithm can still produce errors, even with perfect priors, when the intended sense of a term is not its most frequent sense and there are few other terms in the text that intersect their senses on that intended sense.

Combining Disambiguation with Prior Probabilities Figure 7 shows the output. This approach should provide the most precise model of the content of a text, if the disambiguator performs well and the priors are correct.

Score	Category Number	Category Name
	AP890101-0001	
9.000000	01155	Asian_country Asian_nation
8.200000	00865	movie film picture
6.000000	01932	army_unit
5.492064	01005	war warfare
4.787696	03805	people
4.146906	03738	happening occurrence natural_event
3.781250	03396	education
3.486759	03584	school
3.369697	01159	American_state
3.000000	00642	right
	AP890109-0002	
3.000000	01158	country state land nation
3.000000	02389	university
3.000000	00178	achromatic_color
2.160050	03861	status social_state
2.105806	02758	citizen
2.000000	00620	calendar_month month
2.000000	01171	council
2.000000	01201	lawyer attorney
2.000000	02782	migrant
2.000000	01189	conservative
	AP890112-0001	
5.686035	02509	court tribunal
4.000000	02758	citizen
3.933333	00773	murder homicide slaying
3.000000	01875	city metropolis urban_center
3.000000	01201	lawyer attorney
2.663432	00620	calendar_month month
2.281482	00353	request petition solicitation
2.000000	02821	general_officer
2.000000	02717	expert
2.000000	01157	North_American_country

Figure 6: Prior Probability Weighted Direct Evidence

In the first text, this is not the case. Movie [865] dropped down to sixth because the term film was disambiguated incorrectly (although a different sort on ties would move it to fourth). On the up side, right [642] moved up to fifth. A second disambiguation error, that of assigning Mississippi the sense river [3146] instead of American_state [1159], adds a bad category to the classification. Coming in tenth is worker [2831] one of the two senses of the term "volunteer," the other being a volunteer in a military context.

The second text is still problematic. Writing [350] is on point, as are perhaps country [1158] and citizen [2758]. And while this is still better than randomly assigning categories, it is not notably better than any of the alternative algorithms.

The final text comes out about as well as with the priors alone, with the eighth and ninth categories suspect. Expert [2717] came from "sniper," which has no other sense in the lexicon. Medical_building [2375] comes from the incorrect disambiguation of "home."

Incorporating the disambiguation information is a two-edged sword. On the one hand it gains the writing category for the second text. On the other it falls down in the first and the third, allowing bad categories to rise in the rankings. Improvements in disambiguation should translate directly into better classifications.

Although it is difficult to distinguish between the direct evidence methods based on their performance here, it is possible to distinguish between those methods and the associational approach. Compared to the direct methods, the associational algorithm produces less satisfying classifications. This is, of course, a very subjective evaluation, hampered by the quality of the category set and the need for too many "judgment calls" when determining whether or not a given categorization is felicitous. I next explore a more quantitative evaluation of algorithm's performance.

Score	Category Number	Category Name
	AP890101-0001	
9.000000	01155	Asian_country Asian_nation
7.933631	03738	happening occurrence natural_event
7.000000	01005	war warfare
6.000000	01932	army_unit
6.000000	00642	right
6.000000	00865	movie film picture
5.722930	03146	river
5.539576	03805	people
5.000000	03584	school
3.569444	02831	worker
	AP890109-0002	
4.521965	01158	country state land nation
3.000000	00350	writing writings written_material
3.000000	02758	citizen
3.000000	00178	achromatic_color
3.000000	02389	university
2.080025	03861	status social_state
2.000000	02729	scholar scholarly_person student
2.000000	00290	language linguistic_communication
2.000000	00620	calendar_month month
2.000000	01189	conservative
	AP890112-0001	
9.000000	02509	court tribunal
4.000000	01157	North_American_country
4.000000	02758	citizen
3.977778	00773	murder homicide slaying
3.627478	02587	government authorities regime
3.000000	00353	request petition solicitation
3.000000	01875	city metropolis urban_center
2.931973	02717	expert
2.000000	02375	medical_building
2.000000	01148	national_capital

Figure 7: Combined Disambiguation and Prior Probabilities

7.3.2 Quantitative Analysis

For the quantitative analysis, I only consider the qualitatively promising algorithms, omitting the unweighted and uniform distribution weighted approaches. Keeping in mind the deficiencies of the thesaurus, I chose six of the more coherent categories that cover topics that a person might actually be interested in. These categories are:

MURDER Humans killing humans with intent.

KILLING Anything killing anything, with or without intent.

DUE-PROCESS Due process of law, legal proceedings, trials.

WARSHIP Military naval vessels and their activities.

FIREWORK Explosives of the Fourth of July type.

GRAIN Corn and wheat as crops and commodities.

For each of these categories I examined every text that had it assigned as on of its top three labels. The rankings of the top three were ignored. This allowed for a simple relevant/irrelevant decision: a text is marked correct if the category assigned to it applies, in the sense that the text would be accepted as relevant when retrieving documents about that category.

Figure 8 shows the precision results for the three algorithms. Precision is the ratio of correctly classified texts to the total number of texts labeled with the categories in question. Looking at the totals, it is apparent that the direct methods outperform the associational approach.

The other dimension for evaluating a text classification is recall, the ratio of the number of correctly classified texts to the number of texts that should have that classification. The AP corpus is not tagged with respect to this category set, but the number of relevant documents can be approximated by combining the lists of texts deemed correct in the evaluation of precision. Unioning the correctly classified texts yields a total of 97 relevant texts. This number can be used to approximate the recall of the three algorithms. The results are shown in Figure 9, along with the precision scores. Because I am approximating recall based on the output from these algorithms, I expect that a number of the relevant texts were missed, making the true recall lower. However, increasing the number of relevant documents would not change the

	Associational		Priors		Disam & Priors	
	Correct	N	Correct	N	Correct	N
MURDER [773]	4	4	37	37	29	29
KILLING [774]	0	0	5	8	9	10
DUE-PROCESS [1053]	7	8	7	8	11	14
WARSHIP [2077]	7	10	5	7	6	8
FIREWORK [2126]	2	5	2	5	2	5
GRAIN [2893]	14	20	18	23	15	17
Total	34	47	74	88	72	83
Precision	72.34		84.09		86.75	

Figure 8: Classification Precision

ordering of the scores. Looking at the scores, the associational method has a much lower recall than either of the direct methods, and does not show the expected recall/precision trade-off. Its recall is less than half of the other two methods, and its precision is much lower.

The trade-off is seen when comparing the use of priors alone to disambiguation plus priors. Because the two algorithms are using almost exactly the same data to compute their labelings, there is little difference in their performance. Using the more precise disambiguation information costs some recall, due in part to errors in the disambiguation, but gains a comparable amount in precision.

All three of these algorithms use the same association training data, execute in about the same amount of time and space, and produce interesting classifications. The associational algorithm does not keep pace with the direct methods. Improvements in the training, using a larger training set with wider coverage, would improve the behavior of the associational algorithm. Those same improvements would also be seen in the improved performance of the disambiguation and the concomitant improvement to the quality of the priors. What the direct methods do not capture, that the associational method sometimes does, are those categories that are related to the text, when there are no terms with those senses in the text.

The category writ [338] in the associational classification of the third text is an example of this. None of the terms in that text directly indicate this

Algorithm	Recall	Precision
Associational	35.05	72.34
Priors	76.29	84.09
Disam & Priors	74.23	86.75

Figure 9: Classification Recall

category, but it does tend to co-occur with legal proceedings type terms, like court, hearing, and lawyer. How to combine the two types of evidence so that this type of category is recognized is another question raised by these experiments.

8 Conclusions and Future Directions

This study presents two alternative methods for automatically classifying unrestricted texts with categories automatically derived from WordNet, an on-line thesaurus. What distinguishes the methods is the type of evidence that they use. The first, associational, uses indirect evidence in the form of term–category co-occurrence information, which has proven useful for term sense disambiguation. The second uses direct evidence in the form of the senses of the terms that appear in a text, optionally enhanced by disambiguating the terms. Both approaches demonstrate potential utility, with the direct methods outperforming the indirect. The experiments also identified a number of issues that should be addressed when using these techniques.

The direct evidence methods that incorporate prior probabilities on the term senses, both with and without disambiguation, outperform the associational approach. Qualitatively, the classifications seem more appropriate on a case by case basis. Quantitatively, the direct methods offer both higher precision and a higher approximated recall. One future direction is to determine a more appropriate method for combining the indirect evidence. The simple approach used here, each term contributing to all of the categories it co-occurs with, could be enhanced to include some disambiguation, both in the training and the subsequent deployment. How to do so is an open question.

Automatically constructing categories from WordNet is problematic. The algorithm used categorizes terms based on the size of the clusters, resulting in an uneven level of distinction. For example, there are categories for a number of different species of mushroom, where mushroom by itself would probably do. Then there is the problem of the "time" categories, where the terms map into senses that aren't appropriate for use as classifications. Additionally, there is no relationship between the WordNet synsets for the different parts of speech, making it difficult to consider other terms besides the nouns for these experiments.

These issues could be addressed by getting a better thesaurus on-line, one that provides a single hierarchy across the parts of speech. The most readily available alternative is the on-line 1911 version of Roget's thesaurus, which has the unified hierarchy, but its age makes it likely to provide poor coverage of present day language. A second alternative is to add some human intervention. Here a person could sit down with the categories and manually move terms/categories about to provide a more sensible, even model. This would address the issue, but would be very tedious. A middle road would be to manually identify the contentless categories (time) and filter them out in the same way that stop-words are filtered out. Even without such human intervention, the categories provide a reasonable set of labels for classifying texts.

All of these algorithms are dependent on the quality of the disambiguation computation to some degree. The associational method uses it directly, the others indirectly in the form of the priors. As I have shown, the quality of the disambiguation is somewhat disappointing compared to its potential as demonstrated by [3]. This is the result of training on too small a corpus. Any further research that uses these algorithms should train on at least as much text as Yarowsky used (approx. 10 million words). Insufficient training data is often a problem for frequency based algorithms and this one is no exception.

A related study of automatic text classification was performed by Larson [7], where he concluded that fully automatic classification may not be possible. It is difficult to compare that study to these results, for a number of reasons. First, he used the Library of Congress Classification (LCC) numbers as the category set, which is much larger (over 100,000) than the category set I derived from WordNet. His classifications used only the title and subject heading fields from a document's MARC record, rather than full texts. Finally, selecting the correct LCC code is a more specific task than that of assigning a set of categories to a text.

This study has demonstrated that, given an on-line thesaurus, it is possible to automatically generate a set of conceptual categories. These categories can then be used to classify general free texts with no human intervention. The resulting classifications are qualitatively pleasing, and demonstrate a reasonable degree of precision. As more and more text comes on line, the task of manually classifying it for later access becomes harder and harder. These algorithms offer an automatic alternative, one that can be used by itself, or as an aid in manual classification.

Acknowledgements

I would like to thank my research advisor, Professor Wilensky, for directing this research and keeping me on track. His insights were instrumental in the design and evaluation of these algorithms. Professor Larson, for being the second reader on this project. And I would also like to thank Marti Hearst for providing code for, and the answers to many questions about, the associational labeling and disambiguation. Her help was invaluable in the execution of this project. This material is based in part upon work supported by the National Science Foundation under Infrastructure Grant No. CDA-8722788. This research was supported by CNRI subcontract M1717, from DARPA prime contract MDA972-92-J1029.

References

[1] Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In the *Proceedings of the Second Conference on Applied Natural Language Processing*. Austin, Texas, 1989.

[2] Fisher, D. and Riloff, E. Applying Statistical Techniques to Small Corpora: Benefitting from a Limited Domain. In the *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 1992.

[3] Gale, W., Church, K., Yarowsky, D. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th meeting of the Association for Computational Linguistics*, 249–256. 1992.

[4] Gale, W., Church, K., Yarowsky, D. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 5-6.415–439. 1992.

[5] Hearst, M. and Schütze, H. Customizing a lexicon to better suit a computational task. In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, 55–69. Columbus Ohio, 1993.

[6] Hearst, M. *Context and Structure in Automated Full-Text Information Access*. Doctoral Dissertation University of California, Berkeley. To appear. 1994

[7] Larson, Ray R. Experiments in Automatic Library of Congress Classification. In the *Journal of the American Society for Information Science*, 43(2):130-148, 1992

[8] Liddy, E.D. and Paik, W. Statistically-Guided Word Sense Disambiguation. In the *Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, 1992.

[9] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. J. Introduction to WordNet: An On-line Lexical Data Base. In the *Journal of Lexicography*, vol 3, no. 4, pp 235-244, 1990.

[10] Yarowsky, D. Word sense disambiguation using statistical models of Roget's categories trained on large corpora. *Proceedings of the Fourteenth International Conference on Computational Linguistics*, pp. 454–460, Nantes, France, 1992.

A Sample AP Texts

A.1 Movie Time AP890101–0001

The celluloid torch has been passed to a new generation: filmmakers who grew up in the 1960s. “Platoon,” “Running on Empty,” “1969” and “Mississippi Burning” are among the movies released in the past two years from writers and directors who brought their own experiences of that turbulent decade to the screen.

“The contemporaries of the ’60s are some of the filmmakers of the ’80s. It’s natural,” said Robert Friedman, the senior vice president of worldwide advertising and publicity at Warner Bros. Chris Gerolmo, who wrote the screenplay for “Mississippi Burning,” noted that the sheer passage of time has allowed him and others to express their feelings about the decade.

“Distance is important,” he said. “I believe there’s a lot of thinking about that time and America in general.” The Vietnam War was a defining experience for many people in the ’60s, shattering the consensus that the United States had a right, even a moral duty to intervene in conflicts around the world. Even today, politicians talk disparagingly of the “Vietnam Syndrome” in referring to the country’s reluctance to use military force to settle disputes.

“I think future historians will talk about Vietnam as one of the near destructions of American society,” said Urie Brofenbrenner, a professor of sociology at Cornell University.

“In World War II, we knew what we were fighting for, but not in Vietnam.”

“Full Metal Jacket,” “Gardens of Stone,” “Platoon,” “Good Morning, Vietnam,” “Hamburger Hill” and “Bat 21” all use the war as a dramatic backdrop and show how it shaped characters’ lives. The Vietnam War has remained an emotional issue in the United States as veterans have struggled to come to terms with their experiences. One was Oliver Stone, who wrote and directed the Academy Award-winning “Platoon.”

“I saw ‘Platoon’ eight times,” said John J. Anderson, a Palm Beach County sheriff’s lieutenant who served in Vietnam in 1966-67. “I cried the first time I saw it ... and the third and fourth times. ‘Platoon’ helped me understand.”

Stone, who based “Platoon” on some of his own experiences as a grunt, said the film brought up issues that had yet to be resolved. “People are responding to the fact that it’s real. They’re curious about the war in Vietnam after 20 years,” he said. While Southeast Asia was the pivotal foreign issue in American society of the ’60s, civil rights was the major domestic issue. The civil rights movement reached its peak in the “Freedom Summer” of 1964, when large groups of volunteers headed South to help register black voters.

In “Five Corners,” a movie about the summer of ’64 in the Bronx starring Jodie Foster, her friend, played by Tim Robbins, leaves his neighborhood to volunteer in the South after seeing the Rev. Martin Luther King Jr. on television.

Alan Parker’s “Mississippi Burning” focuses on an incident that clouded the Mississippi Summer Project — when 1,000 young volunteers from mainstream America swept into the state to help register black voters. The movie is a fictionalized account of the disappearance and slaying of three civil rights workers: Michael Schwerner, Andrew Goodman and James Chaney.

They were reported missing on June 21, several hours after being stopped for speeding near Philadelphia, Miss. After a nationally publicized search, their bodies were discovered Aug. 4 on a farm just outside the town.

One of those who recalled the incident was Gerolmo, a student in the New York public school system at the time. The screenwriter said the incident had a powerful effect on his way of thinking. “It was the first time I ever considered that our country could be wrong,” Gerolmo said.

The film stars Willem Dafoe and Gene Hackman star as FBI agents who try to find the bodies of the missing workers and overcome fierce local resistance to solve the crime.

In a more offbeat and outrageous way, John Waters’ “Hairspray” discusses integration in Baltimore in 1963 when a group of teen-agers tries to break down the barriers of a segregated dance show.

Also set in Baltimore is Barry Levinson’s “Tin Men,” starring Danny DeVito and Richard Dreyfuss as two slick aluminum siding salesmen in the early ’60s. The movie mirrored a squarely middle-class culture, one that was not caught up in sex, politics and drugs.

Instead of focusing on a well-known historic event, writer-director Ernest Thompson takes a more personal approach in “1969.” Robert Downey Jr. and Keifer Sutherland star as college students who battle their parents and each other over sex, drugs and the Vietnam War.

“I was 19 in 1969. It was a fulcrum time for me,” said Thompson, who was a student at American University at the time. “I think it was just the right time in my growth as an artist and as a man to try to write about something that happened in my youth.” “Running on Empty” takes place in the ’80s but the ’60s are much in evidence. Judd Hirsch and Christine Lahti play anti-war activists who sabatoged a napalm plant in 1970 and are forced to live underground with their two children.

Naomi Foner, who wrote “Running on Empty” and also served as the film’s executive producer, grew up in Brooklyn, N.Y., the daughter of sociologists. Her own experiences made Foner well qualified to give “Running on Empty” its strong political theme. "I lived through that time and I've wanted to find the right way to present it to this generation," said Foner, a member of the radical Students for a Democratic Society while attending graduate school at Columbia University.

Foner, who also taught in Harlem's Head Start program and helped register voters in South Carolina, said many young people are curious about what happened in the '60s.

"A lot of them think it was an exciting time that they were sorry to have missed," she said. Brofenbrenner said movies are a good indicator of the concerns of the general public: "The principle impact of the media is that they reflect the values of the larger society.

"Film is a very powerful art medium," he said. "I believe it very accurately reflects not only the prevailing but the coming trends. It's because film writers, like other writers, are perceptive people. They get the message of what's going on."

A.2 Immigration AP890109-0002

A former Yale University lecturer who was stripped of his American citizenship in 1986 for his role as a Nazi propagandist in the Soviet Union during World War II has asked for refugee status in Canada, a report said Monday.

Vladimir Sokolov disappeared in July when he was scheduled to appear at a deportation hearing in Hartford, Conn. His whereabouts were unknown until he applied for refugee status in Montreal sometime before Jan. 1, claiming that his life would be in danger if he was forced to return to the Soviet Union, the Canadian Broadcast Corp. reported.

No date has been set for an immigration hearing, the report said. From 1942 to 1944, Sokolov was a writer and editor of a Russian language newspaper published by the German army in his hometown of Orel, 220 miles south of Moscow. Anti-Semitic articles appeared under his name, although he has maintained that the most offensive tracts were written by Nazi censors.

His Canadian lawyer, Julius Grey, said Sokolov faces almost certain death if deported to the Soviet Union.

"The Soviet press has been gloating over his return and have called him a traitor," Grey said. "He would likely be put to death or given a lengthy sentence. For all practical purposes it would be the end of his life."

Grey, a noted constitutional lawyer who is also defending convicted murderer John Joseph Kindler from extradition to the United States where he faces the death penalty, said the 75-year-old Sokolov is in very poor health.

Sokolov failed to reveal his wartime activities in 1951 when he entered the United States as a displaced person. He became a citizen in 1957 and two years later began lecturing on Russian language and Soviet dissident literature at Yale.

His past was uncovered by the Yale student paper in 1976 and he later resigned. But the U.S. government waited until 1982 before filing a complaint to strip him of his American citizenship.

A.3 Deportation AP890112-0001

A Salvadoran accused of conspiring in the 1980 assassination of Archbishop Oscar Romero was freed on bail Thursday night after his homeland's high court denied a request for his extradition, his attorney said.

Alvaro Saravia had been ordered held by President Jose Napoleon Duarte's government, which accused him of arranging Romero's assassination on orders of right-wing legislator Roberto d'Aubuisson, who denied the allegation.

"Saravia telephoned one of my associates at 8:30 tonight and told him 'I'm out,"' said his Miami attorney, Neal Sonnett. Saravia was released on $10,000 bond pending deportation hearings, he said. The Salvadoran Supreme Court ruled last month there wasn't enough evidence linking Saravia to the murder, and said he would not be subject to arrest if he were sent home.

Saravia, a former captain in El Salvador's air force, had been jailed in Miami since November 1987, when he was arrested for a traffic violation and authorities found he had been in the country illegally since 1985.

The Salvadoran government withdrew its extradition request after its courts ruled that there were no grounds for arresting him on charges he had violated his visa, and that the extradition request itself was illegal.

Romero, an outspoken critic of right-wing death squads, was shot by a sniper while saying Mass in San Salvador on March 24, 1980. In November 1987, President Jose Napoleon Duarte's government accused Saravia of arranging the assassination and released the testimony of Amadeo Garay Reyes, who allegedly drove the sniper to the church.

The Supreme Court ruled that Garay's testimony was not credible, partly because he waited more than seven years to come forward with his story.

The court also said Attorney General Giron Flores did not have the constitutional power to ask the United States to send Saravia back to El Salvador.

Flores had directly asked the United States for Saravia's extradition. The case went to the Supreme Court after Saravia's attorney requested a hearing.

Sonnett, who is also representing Panamanian Gen. Manuel Antonio Noriega against federal drug charges here, said Saravia had gone home to his wife, who has been living in Miami.