Using Generic Corpora to Learn Domain-Specific Terminology

David Vogel
The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
1-703-883-6238
dvogel@mitre.org

ABSTRACT
This paper describes a knowledge-weak technique for automatically learning terminology relevant to a given domain from a corpus of domain-specific documents. We used a generic corpus as a filter for scoring the relevance of terms to a domain. We tested this approach against three corpora from different domains and, in each case, high-scoring terms consistently represented concepts relevant to the domain from which they came.

Categories and Subject Descriptors
H.4.3 [**Communications Applications**]: Information Browsers.

General Terms
Algorithms, Experimentation, Languages.

Keywords
Term Discovery, Conceptual Browsing, Machine Learning, Information Extraction, Text Processing, Log Likelihood Ratio.

1. INTRODUCTION
The comedian Steve Martin gave his sure-fire scheme for getting rich in a routine that began, "Okay, first you get a million dollars. Then…." The joke is that the goal is also a precondition. There is an analogous paradox in the field of machine learning: knowledge is a precondition for learning. The knowledge does not have to reside in the entity doing the learning. In the case of induction algorithms, such as decision-trees and neural nets, which learn from positive and negative examples, the knowledge is present in the human who selects the examples. But there has to be pre-existing knowledge somewhere in the system.

This paper describes work done for the Conceptual Browsing MITRE reseach project. The goal of Conceptual Browsing is to develop software that automatically constructs Yahoo-like ontologies from domain-specific text collections. There is an obvious need for such a capability. Insert your favorite information glut factoid here; mine is: the amount of information generated each year is between one and two exabytes[1] [22]. It is not possible to manually profile the contents of every on-line text repository and then organize the profiles into a browseable data structure.

The capability to automatically construct ontologies of text databases would have several important applications. One would be to fill in the gaps in manually constructed ontologies, another to develop sophisticated profiles of text repositories for next-generation Internet resource discovery utilities. An automatic ontology builder could also serve as the front end for data mining applications, turning raw text into structured fodder they can handle. Of course, an automatic ontology builder would be a data mining application in its own right.

This paper describes an approach to satisfy the necessary first step in automatic ontology building. Terminology determines ontology: before one can discover the relationships between discrete bits of knowledge, one needs to find the bits, the vocabulary of the domain.

Our approach to discovering terminology in raw text term relies on a new idea, or at least an idea new to the field of information extraction. We use a generic corpus as a background against which to highlight the prominent terms in a domain-specific corpus. The idea comes from text categorization, where feature selection techniques are used to find the best features, words, with which to distinguish among documents on different topics. Surprisingly, very little work is required to compile an adequate background corpus. In fact, the whole approach is knowledge weak. We rely on statistical analysis of term occurrences to point out the meaningful terms.

There is no getting away from the fact that some knowledge is necessary to learning anything new. But the less up front knowledge required, the better. Knowledge is costly, in time, human effort and, as a result, money. Systems dependent on knowledge tend to be brittle, requiring extensive overhauling when they are applied to new domains and languages. Our goal is to maximize the quality of extracted terms while minimizing the cost of finding them.

Based on the experimental results reported below, we have made promising progress toward our goal. We used generic filters to

[1] An exabyte is 10^{18} bytes.

extract terminology from three corpora from the following domains: electric automobiles, federal income tax, and infectious diseases.

When we use the word term in the paper, we mean one or more words considered as unit. For example, *ford motor company* is a trigram that is a single term. The bigram *ford motor* is also a term, as are the unigrams *ford*, *motor*, and *company*. When necessary, we specify the type of term we mean. The word *word* always denotes a single word.

Section 2.0 describes related work and Section 3.0 our approach. The characteristics of the target and background data are discussed in Section 4.0, as well as the preparation of the data for analysis. Section 5.0 compares term scoring statistics, and we present the results of term scoring experiments in Section 6.0. Conclusions and possible future directions are discussed in Section 7.0.

2. RELATED WORK

Most approaches to automatically extracting information from text collections require some form of knowledge to get started. Any system that parses text or just tags parts of speech, must know, to some degree, the syntax of the language it is parsing. Systems that automatically build thesauri by extracting meaningful term collocations—i.e., consecutive strings of words that re-occur in a corpus, indicating a semantic basis got their cohesiveness—typically depend on syntactic knowledge. SEXTANT [18] and Xtract [34] are two early examples. More recently, Boeing's Expert Locator began with an extensive domain-specific thesaurus, built up over many years by company librarians [6]. Initially, the researchers manually constructed a conceptual index of expertise within Boeing by mapping experts' words about their own knowledge into the thesaurus. New concepts and relations were automatically added to thesaurus using a variety of techniques, including computing subsumption of compound words [35], and inferring relations between noun-noun and adjective-noun phrases based on NLP and concept heuristics. MindNet, developed by Microsoft researchers, uses NLP-based tools for automatically extracting semantic information from definitions and example sentences in highly formatted on-line knowledge sources: machine-readable dictionaries and Microsoft's Encarta 98 Encyclopedia [32]. MITRE researchers developed a system for extracting proper names from newswire [24]. Using knowledge of syntax, honorifics, and discourse, the system recognizes previously unknown names. The Snowball system learns to extract information in structured text [1]. Given initial training examples of relevant structural pattern, Snowball is able to discover new patterns on its own. The Knowledge Acquisition from Text (KAT) system also starts with seed concepts to then automatically discover new concepts and relations in text on the World Wide Web [28]. CMU researchers have developed a trainable information extraction system that uses an ontology, and training examples of lexical items of interest, to automatically build and maintain knowledge bases with information learned from the Web [10].

Information retrieval researchers have been trying to improve search engines by incorporating knowledge into the indexing process. CLARIT [14] added to the standard statistical indexing methods, natural language processing techniques and world knowledge in the form of a domain-specific thesaurus, to better identify terms for indexing documents. In addition to linguistic and world knowledge, Woods [35] employed a sophisticated subsumption-based knowledge representation scheme to build a conceptual taxonomy of a document collection.

A more prosaic strategy to increase the ability of search engines to retrieve relevant documents is to limit the documents indexed by a search engine to a particular domain. There are many examples of domain-specific search engines, including the CACTVS Chemistry Spider for chemical databases, MathSearch for mathematics and statistics documents, and Social Science Information Gateway for resources in the social sciences [20].

There have been efforts to use knowledge-free techniques, beyond the traditional information retrieval algorithms, to analyze text corpora. Scatter/Gather is an intertactive tool that helps users browse through large collections of documents [11]. The system automatically groups a corpus into disjoint subsets, the Scatter phase, by creating word vectors representing cluster centroids and then assigning documents to one of the centroids based on nearest-neighbor analysis, the Gather phase. A summary of the contents of each cluster is culled from the most central documents, those closest to the centroid, and the most central words, those that appear most frequently throughout the group. No claim is made that this approach extracts highly accurate models of the clusters. The cluster summaries provide the user with hints of the contents of the corpus, and the user can recluster the corpus using words from the summaries as seeds for Scatter. GlOSS is a resource discovery utility that maintains a metadatabase of statistical models of text databases [17]. The simple models are based only on word counts, but nevertheless enable GlOSS to direct search engines to good sources of information for a given query. Callan [4] describes another knowledge-weak approach for profiling the contents of text databases: using machine-generated queries to retrieve a representative sample of documents. The initial query term is selected randomly from the TREC corpus. The rest of the queries are generated by continuing to select terms from the TREC corpus, or by choosing terms from the retrieved documents. Experiments indicate 100 single-term queries retrieve representative samples of large text databases. Simple language models can be created from these samples for use by resource discovery utilities.

The jumping off point for the Conceptual Browsing project was [33]. University of Massachusetts researchers bootstrapped ontologies from the information in small corpora using limited pre-existing knowledge and no training data. Their technique depends on the corpus being monothetic; i.e., about one topic. Each monothetic corpus was created from TREC documents retrieved with a single query; e.g, the TREC Topic 230 query: "Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?" The query is, of course, an important source of knowledge. Selection of terms from which to build the hierarchy is based several criteria: does a term appear in the query, or in an expansion of the query using local context analysis; does a term appear more frequently in the best passage of documents relevant to the query than non-relevant documents. The best passage is the one most similar to the query. They then use context-based subsumption to build a hierarchy of relations from the base terms.

The Conceptual Browsing project is extending this approach to bootstrapping an ontology from monothetic corpora. We now describe a new knowledge-weak approach for performing the first step of that process, extracting relevant terminology from domain-specific document collections.

3. APPROACH

Our approach learns terminology contained in a corpus without making any demands on humans beyond ensuring the documents are from a single domain. A person does not need to supply an initial query to create the corpus, training examples, or any other knowledge about the corpus in particular or the domain discussed in the documents. We use pre-existing generic collections of documents as the sole basis for recognizing terminology in domain-specific corpora.

The approach is based on a standard feature selection technique used to train categorizers. Feature selection refers to the process of discovering the best set of features for distinguishing among members of different categories [3]. The white and gray ovals in Figure 1 represent the sets of features possessed by two categories, the black oval their intersection. A feature selection algorithm ignores common features, instead determining the best subset from the white and gray regions to use for deciding category membership. Typically considerations besides limiting the number of categorization errors come into play when selecting features, such as cost, generality and simplicity.

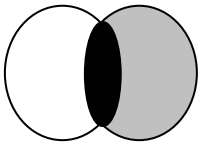

Figure 1. Features of two categories.

Selecting the most representative features for a single category is similar but different from categorization. In the case of categorization, if a feature, say F1, always appears in members of one category, C1, and never appears in members of C2, then the feature is perfect for categorizing members of both categories. If an entity has F1, it is in C1; otherwise, it is in C2. In the case of finding the most representative features, i.e. terms, in a domain-specific corpus, we have no interest in terms that never, or infrequently appear in the target. Figure 2 represents terms in a target corpus as white oval and those in a background corpus as a gray oval. We are interested only in the part of the white oval not overlapped by the gray. The gray oval filters out common terms.

To learn the most representative terms of a domain-specific corpus, its terminology, we compare the distribution of terms within this target corpus and in a more general, polythetic corpus. Using the general corpus as a background filter, we expected the domain-specific terms in the target corpus to stand out as prominent features.

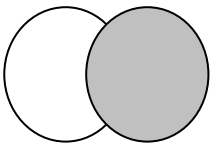

Figure 2. Background filters out common terms.

Not a lot of work went into selecting the background corpora that produced the promising results presented below. We used two background corpora for the research reported here. We performed no special analysis on either of the background corpora we created to measure their generality, aside from observing that they contained documents about more different types of topics than the targets. Other obvious sources for generic corpora include newspaper articles and encyclopedias.

We investigated the potential of background filters for extracting terminology by comparing them against documents from three domains: automobile technology, the Topic 230 corpus; federal income tax, IRS Publication 17; and infectious disease, ProMed.

4. DATA

The three domain-specific corpora used in this research were called: Topic 230, IRS Publication 17, and ProMed. Topic 230 consists of documents relevant to this TREC query: “Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?” IRS Publication 17 is an on-line Internal Revenue Service document that provides general information for filing federal tax returns [19]. ProMed is an E-mail mailing list devoted to outbreaks of infectious diseases and toxins around the world [31].

The two background corpora are called Not Topic 230 and Reuters. The Not corpus consists of a subset of TREC documents not relevant to Topic 230. The Reuters corpus is a subset of the standard research corpus Reuters-21578 [21].

The contents and characteristics of all these corpora are described in detail below. First we show how raw text was transformed into the data from which terminology was extracted.

4.1 Data Preparation

Prior to compiling term statistics, the raw text from target and background corpora underwent a series of transformations. First, stop words were replaced with a no-op token <X>. We did not merely delete stop words because if they were removed without a trace, the transformation would produce bogus co-occurrences. If *the* were removed from the phrase *kick the ball*, then down-stream processing that computed bigrams and trigrams would incorrectly find the bigram *kick ball*. If we were purer of heart, we could have eschewed the use of stop words, since they are an instance of the type of prior knowledge we want to avoid. However, this prior knowledge required no work on our part; the list of stop words was compiled for a previous project, so we used it.

One aspect of data preparation did require work on our part. For each corpus, we had to write software to strip from documents meta-information unrelated to their content. We had to strip headers from the e-mails in the ProMed corpus, HTML tags from the Reuters corpus, and, in general, reformat all documents into a canonical form. Otherwise, our learning technique will find e-mail routing paths and document formatting instructions to be interesting domain-specific terminology. The amount of work required to reformat new corpora will lessen over time as existing filters are re-used against new corpora with familiar formats.

We ignore punctuation, including sentence boundaries, but recognize paragraphs as distinct contexts within a document. Ignoring sentence boundaries caused us some problems. Given the end and beginning of these two sentences “...the car had a Ford

motor. Company officials said...” our downstream processing thinks the trigram *ford motor company* occurred in the original text. Knowledge was the evil needed to correct this problem, in the form of software that recognized sentence boundaries or could parse text.

Hyphenated words were dehyphenated according to these two rules: (1) If both elements of a hyphenated pair are in the dictionary [8] then separate them; e.g., sports-car -> sports car. (2) If a hyphenated pair is separated by a newline, join them unless the hyphenated version is in the dictionary; e.g., sports-\ncar -> sportscar.

Words were stemmed with the Porter stemmer [30]. Each stem was further into a lemma representing the equivalence class of words mapped to the same stem. The longest of the stemmed words from each class was chosen as the lemma. For example, *announcement* was chosen as the representative for all words mapped to the stem *announc*: announcing, announcement, announced, announcements, announce. Stems had to be remapped to actuals words because they were to be used to index documents by an the IR system referred to in the introduction. A user querying the Conceptual Browsing retrieval engine for documents having something to do with, say, electricity might include in a query the term *electricity* or *electric* or some other variant meaningful to humans, but not the stem *electr*.

The last preparatory step was to create compound words from separate words. If following the previous steps the terms *automaker* and *auto maker* appeared, we replaced every occurrence of *auto maker* with *automaker* if the unigram had a term selection score higher score than the bigram. If not, we did nothing. We did not split unigrams into bigrams.

The transformations radically simplify raw text. Figures 3 and 4 give before and after shots of the beginning of a Topic 230 document.

> General Motors Corp. unveiled a prototype electric car it says outpaces some gas-burning sports cars and runs twice as far between charges than previous electric models. The two-seater Impact, which tapers at the rear like a Citroen, can travel 120 miles at 55 mph before recharging and zooms from 0 to 60 mph in eight seconds, GM Chairman Roger Smith said at a news conference Wednesday.

Figure 3. Before.

All transformations were done on the target corpus first. Decisions made for the target—which words to dehyphenate, which word represents its stem class, etc.—were automatically followed in preparing the background. The same raw background corpus could be transformed into different prepared backgrounds, depending on the target it was used against.

> generic motor corp unveiled <X> prototype electric car <X> <X> outpaces <X> gas burn sport car <X> run <X> <X> <X> <X> charge <X> <X> electric model <P> <X> two-seater impact <X> taper <X> <X> rear <X> <X> citroen <X> travel <X> mile <X> <X> mph <X> recharge <X> zoom <X> <X> <X> mph <X> <X> <X> <X> chairman roger smith <X> <X> <X> new conference <X>

Figure 4. After.

Note the mapping of *general* to *generic*. The latter is certainly not an ideal canonical form for representing generality. Correcting this would have required additional knowledge.

4.2 Corpora

The target corpora came in three sizes: small, 85 documents, approximately 0.6 megabytes (MB) in raw form; medium, 285 documents, 2.18 MB; and large, 58.85 MB. The small corpus consisted of the 85 TREC documents, 84,080 words, relevant to the TREC Topic 230 about cars and the environment. The documents have to do with the development of electric cars and related technologies by the big U.S. automobile companies. The medium corpus is a single document from the Internal Revenue Service (IRS) with a lot of general information for taxpayers called *IRS Publication 17*. It consists of 285 short chapters, 258,904 words. The large target corpus contains 11,198 e-mail postings, 4,548,084 words, from ProMed, a mailing list concerned with infectious disease outbreaks and toxins.

Table 1. Corpora Statistics

Corpus	Bytes	Word Types	Entropy
Topic 230 raw	604,501	8,467	10.14
prepared	481,650	4,536	5.54
IRS Pub 17 raw	2,181,939	6,250	9.22
prepared	1,406,579	2,911	4.50
ProMed raw	58,849,844	106,849	11.16
prepared	25,438,196	64,230	5.86
Not 230 raw	4,750,247	39,975	11.33
prepared	3,610,597	20,729	6.23
Reuters raw	27,636,766	68,466	10.64
prepared	15,530,776	31,226	5.99

The background corpus called Not Topic 230 contains 1,105 documents, 631,443 words, randomly selected from four subsets of TREC data: ap900104, fr880929, sjm_001 and zf2_032. As one might expect, Not Topic 230 has no documents relevant to Topic 230 and was the background used to extract terminology from that corpus. A subset of articles from Reuters-21578 was the background corpus used to extract terminology from the IRS Pub 17 and ProMed targets. Reuters-21578 is a standard research corpus for the text-processing community [21]. The corpus contains 21,578 documents, though only 19,043 are complete articles. We included only complete Reuters articles, 2,796,354 words, in our background corpus.

Table 1 lists the vital statistics of the raw and prepared versions of the target and background corpora. Target corpora are in italics. The number of bytes and word types and the entropy scores all fell substantially during the transformation of raw to prepared text because of stemming, mapping of stop words to a single type, <X>, and elimination of most meta-information.

Entropy, the corpus statistic listed in the last column of Table 1, is a standard measure of heterogeneity. The formula for entropy is given in Figure 5 [9]. It is based on the probability distribution of

words in the corpus, where p(w) is the frequency of a given word type. The entropy reported in Table 3 is actually per-word entropy, the entropy divided by the total number of words in the corpus. Topic 230 has a per-word entropy of 5.54 bits; the entropy of IRS Pub 17 is less than half that, indicating Pub 17 is more homogeneous than the Topic 230 corpus. Note that the background corpora have higher entropies than the domain-specific corpora, which is what we expected.

$$-\sum_{w} p(w)\log_2 p(w)$$

Figure 5. Entropy.

5. TERM-SCORING STATISTIC

Our technique learns domain-specific terminology by comparing the distribution of terms in a target corpus to their distribution in a background corpus. The presumption is this: if a term occurs significantly more often in a corpus about one topic than in a corpus about many topics, then the term is of some importance to the topic.

The statistic we chose to score target terms is the log likelihood ratio. Table 2 is a two-by-two contingency table. It displays the document frequency (DF) distribution of the term *electric* in two corpora, the target Topic 230 and the background Not Topic 230 with a total of 1,190 documents between them. DF is the number of documents in a corpus in which a term appears. *electric* appeared in 80 of 85 Topic 230 documents and 61 of 1,105 documents in Not Topic 230. not *electric* refers to the number of documents in which *electric* does not appear.

Table 2. *electric* DF Contingency Table

	Topic 230	**Not 230**	**row**
electric	80	61	141
not *electric*	5	1044	1049
column	85	1105	1190

We used document frequency instead of term frequency (TF), the number of time a term appears throughout a corpus, as the basis for counting term occurrences because DF is less biased. A term appearing many times in a single document is not a good indication of its importance in the corpus as a whole. Here is one example of aberrant behavior by TF. The term *osha* is among the top 10 terms of Topic 230 according to TF. The reason: acronym for the Occupational and Safety Health Organization, occurs 249 times in one Federal Register document. A term appearing in many documents, even if it is just once per document, is a better indication of importance [36].

We compiled contingency tables for unigrams, bigrams, and trigrams in the transformed target and background corpora. The next section describes how we determined when the difference between a term's target and background distributions was significant, indicating it was domain-specific terminology.

What properties were we looking for in a statistic to score the importance of a term in a collection of documents? We wanted a measure that heavily weighted terms of interest, even if they do not occur very often in the target corpus. We also wanted to limit false positives that would overestimate the importance of frequently occurring terms. We compared several candidate term-scoring statistics. Table 6 compares the top 10 scoring unigrams from Topic 230 according to three statistics: mutual information (MI), information gain (IG), and log likelihood ratio (LLR). Table entries in bold are those we judged as not important to the topic of developing electric automobiles.

Table 6. Top 10 Topic 230 Unigrams

MI	**IG**	**LLR**
corporation	electric	electric
announcement	development	development
operator	car	car
regulation	vehicle	vehicle
performance	corporation	corporation
emission	motor	motor
university	battery	battery
spending	automaker	automaker
resources	**announcement**	**announcement**
unveiled	ford	ford

We chose the log likelihood ratio statistic for measuring the correlation of terms with a target corpus. IG performed equally well, but we chose LLR for three reasons. One, unlike IG it does not assume a normal distribution of term occurrence. This is also the reason we did not consider Chi-square. Two, we were nonetheless able to use Chi-square score distributions to measure the significance of LLR scores. The LLR scores can be mapped to a Chi-square distribution. A two-by-two contingency table has one degree of freedom [2], which means again Chi-square score above 3.84 is significant, with a 5 percent chance of a Type I error. See [13] for more on how LLR scores map to Chi-square distributions, and for more on the benefits of using LLR in corpus-based statistical analysis. Three, we found from our experiments that LLR had a greater ability to differentiate the importance in a domain of one term from another. For example, MI produced 62 different scores for Topic 230 unigrams compared to 438 for LLR. Many more terms shared the same MI score. The better resolution from LLR is probably because it uses a more sophisticated model of term occurrence.

Figure 6 shows the LLR formula for a two-by-two contingency table. The formula measures the extent to which a hypothesized model of the distribution of cell counts, H_a, differs from the null hypothesis, H_o. A model's score is its maximum likelihood estimate; i.e., how well the model predicts the actual counts. Note that the null hypothesis has one less parameter than the alternative hypothesis. That is because in our null hypothesis the distribution of a term is the same in the background and target corpus. We can use a single percentage for both. The other parameters are the actual document frequency counts of a term in the two corpora, k1

and k2, and the sizes of the two corpora, n1 and n2. The alternative hypothesis uses the actual document frequencies from each corpus to estimate the percentage of documents in containing the term. The actual model used for H_o and H_a can be whatever you wish. We used a binomial model.

$$-2\log_2\left(\frac{H_o(p;k1,n1,k2,n2)}{H_a(p1,p2;k1,n1,k2,n2)}\right)$$

Figure 6. Log Likelihood Ratio.

Figure 7 fills in the LLR parameters for *electric* from Topic 230. The null hypothesis is that the percentage of documents containing this term is the same in both corpora. The best estimator of this percentage is the total number of documents the term appears in: 141 of 1,190. Just eyeballing the other parameters in Figure 7, you can see the null hypothesis is way off. *electric* appears in most, 80 of 85, Topic 230 documents, and few, 61 of 1,105 Background documents. The best estimates of our two percentage parameters for our alternative hypothesis are 80/85, 94 percent, and 61/1,105, 5.5 percent.

$$-2\log_2\left(\frac{H_o(0.118;80,85,61,1105)}{H_a(0.941,0.055;80,85,61,1105)}\right)$$

Figure 7. *electric* LLR Parameters.

One negative characteristic of LLR is that it scores highly terms whose distribution in the target corpus differs significantly from the null hypothesis, whether this is due to the term occurring significantly more often in the target or background corpus. As a result, some background terms score above the significance threshold. We tried different null hypotheses to eliminate this characteristic, but the changes did more harm than good. Luckily, the spurious terms are easily recognized and eliminated; they appear in a higher proportion of documents in the background corpus than in the target.

Table 7. Some Topic 230 Term Rankings

Terms	LLR	IG	MI	DF
electric	99.9	99.9	81.3	99.9
car	99.6	99.3	81.5	99.8
battery	99.0	98.2	86.9	98.7
emission	96.5	96.8	99.2	79.1
year	67.9	67.6	25.0	99.2
informal	66.2	66.3	0.2	48.6
record	15.2	15.7	4.4	50.2
osha	0.0	0.0	0.0	0.0

Table 7 lists a few more examples of how the statistics under consideration performed on a small representative sample of interesting and uninteresting unigrams from Topic 230. Terms in bold are ones we judged interesting. We wanted them to rank high and the uninteresting terms, normal text, to rank low. The numbers are percentile rankings. Note that document frequency, used as the basis for computing the other statistics, does quite well on its own distinguishing the wheat from the chaff in this small sample.

6. RESULTS

This section presents the results of using generic background filters to extract terminology from domain-specific corpora. We extracted terms from Topic 230 using the previously described background composed of non-Topic 230 TREC documents. We extracted terms from IRS Pub 17 and ProMed using a background corpus based on the corpus known as Reuters-21578.

We extracted unigrams, bigrams, and trigrams. We only considered unigrams that appeared in at least five documents in the target corpus. This means the minimum document frequency count for the longer ngrams also had to be five. In the future, we would not prevent infrequently occurring unigrams from playing a role in the search for meaningful bigrams and trigrams. Results below show that trigrams that occur significantly more often in the target corpus, even if their document frequency count is low, are usually of interest.

Recall that in some cases the log likelihood ratio gives high scores to terms more prominent in the background than the target corpus. We eliminated these terms from our domain terminology sets. A large percentage of terms scoring above the significant threshold were filtered out from the Topic 230 and IRS Publication 17 corpora. Relatively few such terms, 982 of 25,680, from the much larger ProMed corpus were eliminated in this way. The 'Out' column in Table 8 gives the number of target ngrams with scores above the significance threshold that were eliminated by this last filter. The In' column gives the final count of terms of interest.

When we refer to a term score, we mean the Chi-square score, for one degree of freedom at the five percent level of Type I error, associated with the log likelihood ratio. For brevity, we say LLR score.

Table 8. Significant N-Gram Scores by Corpus

Corpus	1-G	2-G	3-G	Out	In
Topic 230	414	82	21	289	228
IRS Pub 17	889	529	110	689	829
ProMed	10,474	12,294	2,912	982	24,698

6.1 Topic 230

According to their LLR scores, 228 terms that appeared in at least five documents in the Topic 230 corpus were interesting. The lowest score among these terms was 14.87, way above the minimum level of significance for one degree of freedom. These terms include unigrams and bigrams and trigrams composed from the 4,536 word types in the prepared text. Table 9 summarizes the distribution of scores above the significance threshold.

The highest scoring term was *electric car* at 382. Recall that Topic 230 consists of TREC documents relevant to the following

query: “Is the automobile industry making an honest effort to develop and produce an electric-powered automobile?”

Table 9. Topic 230: Significant Scores Distribution

Statistic	Score
Count	228
Mean	51.30
Standard deviation	63.68
Range	14.87 to 381.92
Quartiles	20.79, 27.89, 44.82

Other high-scoring terms seemed to us to be relevant, and many low-scoring, but still significant, terms seemed relevant, too. Table 9 summarizes the scoring statistics for the 228 terms. Few terms scored as highly as *electric car*; the mean LLR score was 51.30 yet 75% of the terms scored below the top quartile cut-off of 44.82. The scores are not normally distributed, so the standard deviation is not too meaningful, but we report it to give some sense of the large variability in LLR scores.

Table 10. Topic 230: Top 5 Terms by Quartile

Quartile	Terms	Scores
1	electric car, electric, development, car, vehicle	382-301
2	mph, fleet, carbon, motor corporation ford, **maintains**	44-40
3	natural gas, toyota, sodium, lansing, gallon	27-26
4	**drain**, toyota motor, subcompact, smog, showroom	20-19

Table 10 gives a brief anecdotal sense of the performance of our technique. The table lists the top five terms from each quartile. For a term to be considered, it had to occur in five or more documents in the target corpus, and not be seen in a higher percentage of background than target documents. This restriction is also true for the other corpora, IRS Publication 17 and ProMed.

All but two of the terms in Table 10 seem relevant to Topic 230’s concern with the commitment of automakers to the development of more environment friendly cars. And four of the five top-scoring terms have to do with cars, electricity, or both, another indication that in some simple yet meaningful way our technique has captured the gestalt of the corpus. Terms that we considered uninteresting, or irrelevant, for the target domain are in bold.

6.2 IRS Publication 17

As one would expect of a Government publication meant to convey important information to the general public, IRS Publication 17 uses a limited vocabulary. It is three times the size of the Topic 230 corpus—in number of documents, bytes and words—but contains only 2,911 word types that appeared in five or more documents, less than two-thirds the number for the smaller corpus. The entropy of Publication 17 is half that of Topic 230. These numbers are for the prepared corpora.

The limited vocabulary of this corpus is composed of a large proportion of terms relevant to its domain. From a base of 2,911 word types, 829 terms—unigrams, bigrams, and trigrams—were discovered to be of interest according to their LLR scores. Table 11 summarizes the distribution of their scores.

Table 11. IRS Pub 17: Significant Scores Distribution

Statistic	Score
Count	829
Mean	121.08
Standard deviation	258.06
Range	32.91 to 2894.41
Quartiles	46.98, 62.65, 110.95

The highest scoring term for Publication 17 was *tax chapter*, with a score of 2894. It leads the terms in Table 12, which shows the top five terms from each score quartile. None of the terms we found to be interest are cause for embarrassment, though we would be happier if *federal income tax*, or any of the trigram’s components, had topped the list instead of *tax chapter*. The term *www*, an acronym for the World Wide Web, is not directly relevant to the tax domain, but is an important concept in a publication that directs taxpayers to important sources of information.

Table 12. IRS Pub 17: Top 5 Terms by Quartile

Quartile	Terms	Scores
1	tax chapter, income tax chapter, federal income tax, federal income, income tax	2894-2548
2	apartment, liable, dental, revenue service, internal revenue service	105-103
3	rollover, mecial care, marriage, insurer proceeds, state law	62
4	www, vocational, rental active, pleasure, medical insurance	45

All of the Quartile 1 terms in Table 12 are components of two phrases: *income tax chapter* and *federal income tax*. Any U.S. taxpayer would agree that two of the terms represent important concepts: *income tax* and *federal income tax*. The word *chapter* included in two of the top five phrases does not; it is an artifact of the way we processed Publication 17, not its domain. The document is composed of 285 sections, 284 chapters and one

index. In order to compute document frequency counts, we considered each section a separate document. The term *chapter* appeared in 284 of the 285 documents in Publication 17, but in only 103 of the 19,043 documents in the background corpus; hence, by our calculations, *chapter* is a very important term in the IRS domain.

Table 13 lists the component terms appearing in the top five terms of Quartile 1. They are ranked in order of our ad hoc estimate of their importance. That is why *tax*, *income*, and *federal* are at the top and *chapter* is at the bottom. Even though our top-ranked terms scored below *chapter*, they still received high scores.

Table 13. Pub 17: Quartile 1 Components

Term	Score
tax	1488
income	1748
federal	1449
income tax	2548
federal income tax	2864
federal income	2839
tax chapter	2877
chapter	2489

We had no practical way to measure the precision of our term rankings. This would have required getting one or more unbiased judges, i.e. not members of the research term, to tag each unigram, bigram, and trigram in IRS Publication 17 with a significance score. The same holds true for recall, though we did find a small lexicon of tax terms against which to compare our term rankings. We used the lexicon to get some sense of how many of the terms we should find significant we in fact did find significant. The lexicon, compiled by the branch of MITRE supporting the IRS, contained 127 terms. Eighty-two of these terms appeared in Publication 17 and, of these, 77 (94%) scored above the significance threshold.

The lowest-scoring unigrams above the significance threshold we found uninteresting., but the lowest-scoring bigrams and trigrams above the threshold were. The bottom five are shown in Table 14.

Table 14. Pub 17: Bottom Significant 2-, 3-Grams

2-Gram	Score	3-Gram	Score
cash payment	18	estate tax deduction	37
pension plan	17	state income tax	36
purchase price	16	life insurance policy	36
short term	7	federal tax law	34
interest payment	6	short term capital	27

6.3 ProMed

There were many interesting terms in ProMed, 24,698 above the significance threshold of 3.84. The distribution of scores was much more skewed than in the other two corpora, Table 15. The highest scores were much higher, but there were relatively few of them. The top five ProMed terms scored over 9,000, but the quartiles of the corpus' scores are all lower than the quartiles of Topic 230 and IRS Pub 17. Half of the scores were below 7.16, and despite the huge range in scores, from near zero to over 14,414, the interquartile range was just above 15.

Table 15. ProMed: Term Statistics

Statistic	Score
Count	24,698
Standard deviation	240.45
Median	7.16
Range	0.45 to 14,414.37
Quartiles	4.47, 7.16, 19.66

Table 16 lists the top five terms from each ProMed score quartile. We are especially gratified that the top-scoring term for the ProMed corpus was *disease*, and that three of the other top five terms embody important concepts in the domain of infectious disase: *health*, *infection*, and *case*. The one fly in the ointment is *mod*. The term *mod* is short for moderator. ProMed is a moderated mailing list, and comments interjected into e-mails by the moderator are tagged with *mod* and his initials. Three of the bolded terms in Table 16 are names: *mackay*, *jeremy*, and *cole*.

Table 17 gives the top five terms from the quartiles of ProMed's 8,045 quartile 1 terms. We consider six terms in this table to be errors: *mod*, *worsen*, *violet*, *benyon*, *barr*, and *augmentation*. The term *worsen* seems relevant to a medical domain, but by itself tells us nothing useful about ProMed. Table 17 has fewer errors than Table 17, where eight of 20 terms are not interesting. This is evidence that the terms in quartile 1 have a higher overall quality than the full set of ProMed terms.

Table 16. ProMed: Top 5 Terms by Quartile

Quartile	Terms	Scores
1	disease, health, case, infection, **mod**	14,414-9,157
2	**mackay**, **jeremy,** flourish, **cole**, antiviral drug	20
3	**virtual**, **sooner**, **chose**, mandatory, skeptic	7
4	guangdong, statistician, **simpler,** raise question, poor country	4

Four of the five terms from the bottom quartile in Table 17 are parts of names: *benyon*, *barr*, *bandundu*, and *babesia*. Two of the names are in bold, *benyon* and barr. They are both parts of names, and by themselves convey nothing interesting about the domain. *benyon* is just part of the name of a ProMed correspondent. The other names are related to important concepts. The unigram *barr*

is associated with two diseases: Guillan-Barre Syndrome and Epstein-Barr Virus. The relatively high score of this term is the product of a number of conflations. First, two different names, Barre and Barr, were naively stemmed to *barr*. Second, the unigram is also part of the name of several ProMed correspondents. As a result the document frequency of *barr* counts unrelated occurrences of a term that means different things in different contexts. The term *bandundu* is the name of a province in Zaire. *babesia* is the genus of a parasite transmitted by ticks that causes a number of infectious diseases [12]. We will have more to say about names.

Table 17. ProMed: Top 5 of Quartile1 Quartiles

Quartile	Terms	Scores
1	disease, health, case, infection, **mod**	14,414-9,157
2	**worsen,** waterborne, **violet,** vertebrate, variant creuztfeldt-jakob	119
3	western blot, urinary, state agriculture, snake, salmonella infection	51
4	**benyon**, **barr**,, bandundu, babesia, **augmentation**	30

Table 18 lists representatives from a yet smaller subset of ProMed terms, the top 1,000. Except for *phone*, all seem very relevant. *pool*, as in pool of standing water, and *dead* bird, as in an indicator of a disease outbreak, appear extensively in ProMed correspondence related to mosquitos. *mosquito* ranked ninety-second with a score of 1744.

Table 18. ProMed: Some Top 1000 Terms

Term	Rank	Score
live	100	1,513
phone	200	979
discover	300	735
nile virus	400	582
vet	500	481
japanese encephalitis	600	409
pathologist	700	347
detail	800	306
pool	900	269
dead bird	1,000	239

Table 19 lists the lowest ranking ProMed bigrams and trigrams above the significance threshold. The bigrams are not interesting, but the trigrams are.

Table 19. ProMed: Bottom Significant 2-, 3-Grams

2-Gram	Score	3-Gram	Score
death loss	4	jeddah saudi arabia	4
dairy industry	4	improvement live standard	4
critic mass	4	beth israel medic	4
author order	4	agriculture ministry spokesman	4
consumer product	4	west african country	4

All the tables in this section attempt to give some indication of the precision of LLR scores. The following graph gives some indication, the best for any of the three target corpora, of the recall performance of our approach to finding domain-specific terminology. Tom McEntee, a member of the MITRE technical staff who is an expert in bacterial agents, manually built an ontology relevant to the ProMed domain. The ontology, built for another project independent of the Conceptual Browsing research effort, contained 3,025 unique terms, of which 1,048 appeared in our ProMed corpus. We ranked the 1,048 terms from highest to lowest score; 606 scored above the significance threshold. The graph in Figure 8 displays the curve for LLR score by rank. Term scores are converted to their base 10 logarithm in order to display the graph in a small space. The smooth descent of the logarithmic scores indicates an exponential relationship between term scores and rank. A few high-ranking terms had very high scores, over 10,000 as we saw earlier, and several hundred terms had fractional LLR scores, below 0 on the log 10 scale.

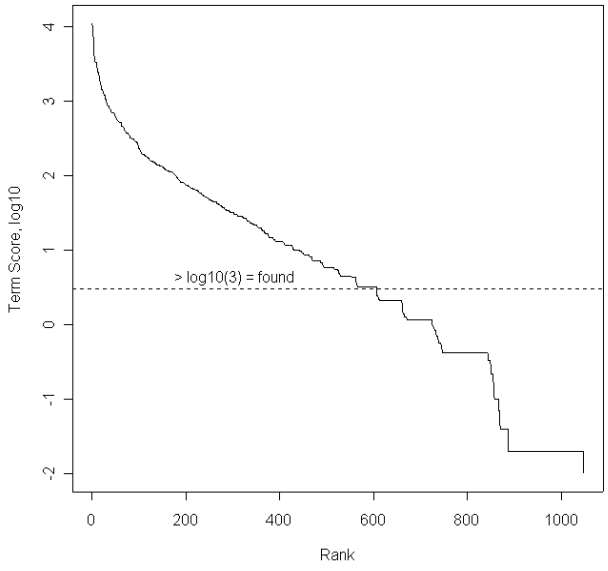

Figure 8. McEntee Term Scores by Rank.

Figure 8 is not precision-recall curve, but something like it. The graph gives some sense of the recall performance in a raw text corpus that has not been tagged with the meta-information

necessary to precisely measure precision and recall. Our approach scored as significant approximately 60 percent of the terms judged important by a domain expert that appeared in the ProMed corpus.

7. CONCLUSIONS AND FUTURE WORK

There is no such thing as a free lunch [15]. However, this paper describes a knowledge-weak technique that costs so little, in terms of human time and effort, that it qualifies as a very cheap lunch indeed when it comes to learning new, non-trivial things about a specific domain.

Using a generic corpus as a filter to find meaningful terminology in a domain-specific corpus appears to be a very good idea. We tried out this idea in three domains of varying size and quality, Topic 230, IRS Pub 17, and ProMed. In each case, the highest scoring terms represented important domain concepts.

Our approach relies on very little knowledge about language, mostly stop words, and none about the domains the documents represent. We require background knowledge in the form of a statistical profile of a generic corpus. The most human-labor intensive knowledge we need is knowledge of the formats of the documents we analyze. We have to understand formats in order to filter out irrelevant metadata. With slightly better knowledge of the format of ProMed correspondence, we could have screened out the annoying *mod* and kept it from appearing in any ProMed results, let alone its top five terms.

Imagine an Internet resource discovery utility that helps search engines find and rank information relevant to a query about a particular topic. Our background filter implementation is probably close to being good enough to provide discovery services with descriptions of domain-specific repositories. Meta-information exactly like that in Table 17, encoded in Resource Description Format (RDF) or some other format understood by Internet utilities, should be of use to Internet search engines.

The work of another group of researchers in knowledge-weak resource discovery suggests a possible synergistic collaboration [4]. They report being able to create a representative sample of a text database using queries generated without knowledge of the contents of the database. If it turns out the sample was collected from a monothetic repository then our knowledge-weak term-extraction technique could be applied to the sample to create a profile the original database. For the sake of this scenario we ignored an important problem, the need to automatically recognize the sample is about one topic. This problem is very similar to a concern of ours: how to measure the generality of a corpus.

The technique described in this paper can be the first step in learning about a new domain. The terms we extract can be the basis for more knowledge-intensive analyses. The Conceptual Browsing project, in fact, takes just such a sequential approach to modeling knowledge in a domain. One module takes in our terms and uses WordNet to discover concepts not explicitly mentioned in the corpus. Another higher-level module takes in terms and looks for relationships among them. This module embodies the ultimate goal of the Conceptual Browsing project, which is to automatically build ontologies of domain-specific corpora.

More can be done to improve the precision of our results without adding knowledge. For example, we probably could reduce the rate of false positives in identifying important domain concepts by using bigrams and trigrams to filter out unigrams whose LLR scores overstate their importance. To consider one possible technique, let us revisit the high scoring ProMed unigrams from Table 17. We said six of the terms, all unigrams, were in error: *mod*, *worsen*, *violet*, *benyon*, *barr*, and *augmentation*. If we add a new filter that requires a unigram to appear in more than one bigram or trigram before qualifying as a term of interest, then five of the six errors are screened out. The only recalcitrant is *mod*.

Table 20 shows the results of the new filter. A unigram gets a plus in the '2-G' column if the term also appears in at least two bigrams with scores above the significance threshold; otherwise, the unigram gets a minus. The same is true in the '3-G' column.

The increase in precision from the Table 20 filters decreases recall. Two important concepts, *snake* and *babesia* are filtered out, though the latter term is part of one high scoring bigram, *babesia microti*. Loss of *snake* is disappointing, since ProMed includes discussion of toxins.

The addition of one type of knowledge would noticeably improve our results: the ability to recognize names, not just of people but also places, organizations and other entities. Of course, such knowledge comes at a cost to the flexibility of our approach. The behavior of names varies among cultures and languages.

The main surprise from our work is that, apparently, background corpora do not have to be carefully chosen. As we said earlier, we expended very little effort creating the two background corpora used to produce the results discussed in this paper. Why is this so? We do not know. Perhaps because natural language is ergodic: a reasonably sized sample is a good model of the whole, and any fairly generic collection of text will do.

Table 20. ProMed: Filtered Concepts

1-Grams	2-G	3-G	Out	In
disease	+	+		disease
health	+	+		health
case	+	+		case
infection	+	+		infection
mod	+	+		**mod**
worsen	-	-	**worsen**	
waterborne	+	-		waterborne
violet	-	-	**violet**	
vertebrate	+	-		vertebrate
urinary	+	+		urinary
snake	-	-	snake	
benyon	-	-	**benyon**	
barr	-	-	**barr**	
bandundu	+	-		bandundu
babesia	-	-	babesia	
augmentation	-	-	**augment ation**	

This is not to say that more work, computational not human, cannot be done to select better background corpora. It remains to be seen whether cheap technology, requiring no or minimal human effort, will suffice. At the moment we have only vague notions of how to measure generality. Much research remains to be done on the analysis and selection of background corpora. A background should be different from a target, but how different? What scales should we use to measure this difference: entropy, number of terms found in a standard reference such as a dictionary or encyclopedia? A background should be about more things than a target, but how many more?

One thing about the nature of a background corpus is obvious. We have to change its mix of documents depending on the type of terminology we want to extract from a target corpus. A background corpus is like a camera filter; the nature of the filter helps determine what you see. Consider a document collection about molecular biology. To extract biological information of interest to the general public, we might select a background corpus composed of articles on many different topics from *The New York Times*. This general background should highlight most of the biology terms in the target, whether or not they were specific to the field of molecular biology. Documents from all fields of biology will use terms such as *cell* and *protein*. Some terms from biology occasionally appear in newspaper articles, but not with the consistency with which they appear in the molecular biology corpus. Of course, when an area of biology becomes news, such as *anthrax* in the months following the September 11 terrorist attacks, then *anthrax* and related terminology is no longer a distinctive feature of domains specializing in biology. Timing is yet another factor to consider in creating a background corpus. Prominent events will skew the composition of newspaper articles.

If we want to use this same target corpus to extract information of interest to a narrower audience, say molecular biologists, then we should try to avoid telling them things they already know; for example, that *cell* is an important term in biology. To avoid discovering useless knowledge, we need to refine the background corpus. If we use a standard biology textbook instead of newspaper articles as our background corpus, then terms specific to the subfield of molecular biology, like *eucaryotic DNA methylase*, should stand out while terms like *cell* are suppressed because they appear with similar regularity in both corpora.

8. ACKNOWLEDGMENTS

Thanks to Inderjeet Mani and Ken Samuel for a stimulating partnership on MITRE's Conceptual Browsing research project.

9. REFERENCES

[1] Agichtein, E., and Gravano, L. Snowball: Extracting Relations from Large Plain-Text Collections in Proceedings of the 5th ACM International Conference on Digital Libraries, 2000.

[2] Agresti, A. An Introduction to Categorical Data Analysis, Wiley, New York, 1996.

[3] Brooks, R. Intelligence Without Reason, MIT AI Lab Memo 1293, April 1991.

[4] Callan, J., Connell, M., and Du, A. Automatic Discovery of Language Models for Text Databases. SIGMOD '99, 479-490, 1999.

[5] Church, K. and Hanks, P. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, 16(1):22-29, 1990.

[6] Clark, P., Thompson, J., Holmback, H., and Duncan, L. Exploiting a Thesaurus-Based Semantic Net for Knowledge-Based Search in Proceedings AAAI-2000, 2000.

[7] Clifton, C., Cooley, R., and Rennie, J. TopCat: Data Mining for Topic Identification in a Text Corpus in Proceedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases, 1999.

[8] COMLEX Syntax Corpus. ftp://cs.nyu.edu/pub/html/comlex.html.

[9] Cover, T. and Thomas, J. Elements of Information Theory, Wiley-Interscience, 1991.

[10] Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. Learning to Extract Symbolic Knowledge from the World Wide Web in Proceedings of AAAI-98, 509-516, 1998.

[11] Cutting, D., Karger, D., Pedersen, J., and Tukey, J. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections in Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992.

[12] Dorland's Illustrated Medical Dictionary. Saunders, 1974.

[13] Dunning, T. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1):61-74, 1993.

[14] Evans, D., et al. Automatic Indexing Using Selective NLP and First-Order Thesauri in Proceedings of RIAO '91, 624-644, 1991.

[15] Friedman, M. There's No Such Thing as a Free Lunch. Open Court Publishing Co., LaSalle, IL, 1975.

[16] Gordon, M., and Lindsay, R. Toward Discovery Support Systems: A Replication, Re-Examination, and Extension of Swanson's Work on Literature-Based Discovery of a Connection Between Raynaud's and Fish Oil. Journal of the American Society for Information Science, 47(2):116-128, 1996.

[17] Gravano, L., Garcia-Molina, H., and Tomasic, A. GlOSS: Text-Source Discovery over the Internet. ACM Transactions on Database Systems, 24(2):229-264, 1999.

[18] Grefenstette, G. Explorations in Automatic Thesaurus Discovery. Kluwer, 1994.

[19] Internal Revenue Service. Publication 17: Your Federal Income Tax. www.irs.gov/forms_pubs/pubs/p17toc.htm, 2000.

[20] Kobayashi, M., and Takeda, K. Information Retrieval on the Web. ACM Computing Surveys, 32(2):144-173, 2001.

[21] Lewis. The Reuters-21578 Text Categorization Test Collection, www.research.att.com/~lewis/reuters21578.html, 1998.

[22] Lyman, P., and Varian, H. How Much Information? www.sims.berkeley.edu/research/projects/how-much-info, 2000.

[23] Manning, C., and Schutze, H. Foundations of Statistical Natual Language Processing. The MIT Press, Cambridge, MA, 1999.

[24] Mani, I., and MacMillan, R. Identifying Unknown Proper Names in Newswire Text in Boguraev, B., Pustejovsky, J. (eds.): Corpus Processing for Lexical Acquisition. MIT Press, 41-59, 1996.

[25] McCallum, A., et al. Building Domain-Specific Search Engines with Machine Learning Techniques in Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace, 1999.

[26] McMahon, J., and Smith, F. A Review of Statistical Language Processing Techniques. Artificial Intelligence Review, 12(5):347-391, 1998.

[27] Mitchell, T. Machine Learning, WCB McGraw-Hill, Boston, 1997.

[28] Moldovan, D., Girju, R., and Rus, R. Domain-Specific Knowledge Acquisition from Text in Proceedings of the Applied Natural Language Processing Conference, Seattle, 2000.

[29] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank Citation Ranking: Bringing Order to the Web in Technical Report available at http://www-db.stanford.edu/~backrub/pageranksub.ps, 1998.

[30] Porter, M. An algorithm for suffix stripping. Program, 14(3):130-137, 1980.

[31] ProMED-Mail. The Global Electronic Reporting System of Emerging Infectious Diseases and Toxins, International Society of Infectious Diseases, 2000. www.promedmail.org/pls/promed/promed.home

[32] Richardson, S., Dolan, W., and Vanderwende, L. MindNet: Acquiring and Structuring Semantic Information from Text in Proceedings of the 17th International Conference on Computational Linguistics, 1098-1102, 1998.

[33] Sanderson, M., and Croft, B. Deriving Concept Hierarchies from Text in Proceedings of the 22nd ACM SIGIR Conference, 206-213, Berkeley, CA, 1999.

[34] Smadja, F. Retrieving Collocations from Text: Xtract. Computational Linguistics, 19(1):143-177, 1993.

[35] Woods, W. Conceptual Indexing: Practical Large-Scale AI for Efficient Information Access in Proceedings AAAI-2000, 1180-1185, 2000.

[36] Yang, Y. and Pedersen, J. A Comparative Study on Feature Selection in Text Categorization in Proceedings of the International Conference on Machine Learning, 1997.