Acquaintance:
Language-Independent Document Categorization by N-Grams

Stephen Huffman
Department of Defense
Ft. George G. Meade, MD 20755-6000

Acquaintance is the name of a novel vector-space n-gram technique for categorizing documents. The technique is completely language-independent, highly garble-resistant, and computationally simple. An unoptimized version of the algorithm was used to process the TREC database in a very short time.

Acquaintance is the name of a technique for information processing that combines the robustness of an n-gram-based algorithm with a novel vector-space model. Acquaintance gauges similarity among documents on the basis of common features, permitting document categorization based on a common language, a common topic, or common subtopics. The algorithm is completely language- and topic- independent, and is resistant to garbling even at the 10% to 15% (character) level. Acquaintance is fully described in Damashek, 1995. The TREC-3 conference provided the first public demonstration and evaluation of this new technique, and TREC-4 provided an opportunity to test its usefulness on several types of text retrieval tasks.

The Acquaintance algorithm can be used for processing sets of documents in two distinct ways. One method explores the conceptual space of a set of documents by determining the degree of similarity among all the documents in that set. When the documents are then viewed with a visualization tool that arranges them so that the distance between them corresponds with their putative degree of similarity, the conceptual space defined by those documents becomes apparent. That is, those documents which are similar, and thus most probably related by language or topic, will cluster together. Furthermore, documents that relate to several different topics will be obvious due to their positions and the strengths of their connections to more than one cluster of documents. Those documents which are not clearly similar to any others in the set will stand alone and unconnected to other documents. This mode of using Acquaintance is very useful when exploring the contents of a large and unknown database, and was used very successfully when applied to the interactive task at TREC-4.

Acquaintance can also be used for the more traditional task of retrieving documents from a database based on specific queries. When used in this manner, reference documents are compared to the documents in the database. Those documents in the database which are similar to the reference documents can be quickly identified. Using Acquaintance in this fashion most closely approximates many of the tasks in TREC, and variations on this latter method were used to process most of the data in TREC-4.

Methodology

N-Gram Processing

The Acquaintance algorithm begins by processing texts in a manner very similar to traditional n-gram based techniques. An n-wide window is stepped through text, moving one character at a time. From each n-gram lying within the window, a hash function generates a value that is treated as an address in a document vector, and the contents of that vector address are incremented by one. When all of the n-grams in the document have been processed, the document vector is normalized by dividing the frequency count of n-grams at each vector address by the total number of n-grams in the document. Thus, the sum of the normalized counts of the n-grams in the document vector will sum to one.

Centroid Subtraction

A crucial aspect of Acquaintance when gauging similarity among documents is the subtraction of a centroid vector from the document vectors. The centroid in Acquaintance defines a context within which a set of documents can be usefully compared. This method of subtracting a centroid stands in contrast to more traditional vector-space models which frequently use some form of multiplicative weighting, which results in a rescaling of the axes in the vector space.

The centroid vector characterizes those features of a set of documents that are more or less common to all the documents, and are therefore of little use in distinguishing among the documents. The Acquaintance centroid thus automatically captures, and mitigates the effect of, those frequent but generally undiagnostic features of the language that are traditionally contained in stop lists and removed by stemming algorithms.

The creation of the centroid vector for a set of documents is straightforward and language independent. After each separate document vector is created, the normalized frequency for each n-gram in that document is added to the corresponding address in a centroid vector. When all documents have been processed, the centroid vector is normalized by dividing the contents of each vector address by the number of documents that the centroid characterizes. A centroid thus represents the “center of mass” of all the document vectors in the set.

Computing Similarity Scores

Once documents are characterized by normalized document vectors, the resulting vector-space model permits the use of geometric techniques to gauge similarity among the documents. When comparing a set of document vectors to a set of reference vectors, the cosine of the angle between each document vector and each reference vector, as viewed from the centroid, is computed using Equation 1:

$$S_{mn} = \frac{\sum_{j=1}^{J} (x_{mj} - \mu_j)(y_{nj} - \mu_j)}{\left[\sum_{j=1}^{J} (x_{mj} - \mu_j)^2 \sum_{j=1}^{J} (y_{nj} - \mu_j)^2 \right]^{1/2}} = \cos\theta_{mn}, \quad m, = 1, \ldots, M, \;\; n = 1, \ldots N \quad (1)$$

where the vectors $x_m,\ m \in 1,\ldots,M$ are the M document vectors, the vectors $y_n,\ n \in 1,\ldots,N$ are the N reference vectors in a J-dimensional space, and μ is the centroid vector.

A cosine value of 1.0 indicates that the document and reference vectors are perfectly correlated (or identical), a value of minus 1.0 that they are perfectly anticorrelated (or antithetical), and a measure of 0.0, that they are uncorrelated (or orthogonal). A great deal of experimentation has been done using this scoring method for gauging topic similarity, and a clear idea of how the measure behaves as features such as n-gram length and garbling are varied (Huffman, in process) has been obtained.

Acquaintance at TREC-4

System Parameters and Text Processing Procedures

In TREC-3, Acquaintance participated for the first time, and was used in just the routing and adhoc tasks. The purpose of participation in TREC-3 was to get a feel for how well a purely statistical system would work compared to more linguistically sophisticated systems. In TREC-4, Acquaintance participated in a much broader range of tasks, including the routing, ad hoc, interactive, filtering, confusion, and Spanish tracks. While the details of the individual tracks will be discussed below, the same software and basic procedure were used in each track.

For the work in TREC-4, a generic, unoptimized version of Acquaintance, written in ANSI C, was used. The TREC data was processed on a heavily time-shared Cray YMP. Both the routing and ad hoc tasks were run as overnight background jobs, and each took less than 8 hours clock time to finish. For most tasks, the n-gram length was five, and the document vector length (or hash table length) was 262144. The only occasions where the n-gram length differed from five was while processing the twenty percent garbled data for the confusion track, when four-grams were used, and for two of the filtering runs, for which seven-grams were used.

Acquaintance requires almost no preprocessing of the documents. To prepare the TREC database, the SGML tags and headers were stripped from the data, and only characters between the TEXT tags were processed. Acquaintance ignored all non-alphabetic characters in the text and translated all lowercase alphabetic characters to uppercase characters.

Routing

The routing task in TREC simulates the process of filtering an incoming stream of documents according to predefined criteria. Participants are given the topic descriptions (which are taken from previous year's TREC conferences) early in the year. However, the the database of documents is not made available until the queries created from the topic descriptions have been formulated and sent into NIST. In addition, and more importantly for Acquaintance, the list of those documents which were judged relevant to each topic is made available to participants. Thus, a large corpus of potential reference documents is available for each routing topic. However, there is no guarantee that the relevant documents from previous years will in fact be representative of the set of documents used as the database in the current year. In TREC-3, the documents used for reference and for the database were very similar. In TREC-4 they were not, and that fact caused problems for Acquaintance.

To perform the routing task, the AP newswire documents from TREC-3 which were defined to be relevant to each of the routing topics were recovered. The goal was to find a useful subset of those documents to use as reference documents against which to compare the documents in the database. To accomplish this, all the supposedly relevant documents for a particular topic were scored against each other, using the Acquaintance metric. Then, that set of documents and associated scores were submitted to the Parentage tool. One feature of this tool, which will be described more fully below, applies graph theory to sets of scored documents to determine which documents in that set are the most highly connected. Taking advantage of this feature, roughly the 50 most highly connected documents for each topic were selected. Those documents constituted the final set of reference documents for that topic. This process thus produced a set of about 2500 reference documents against which the documents in the database were measured for similarity.

To find relevant documents in the database, a document vector from each document was created and the cosine of the angle between that document vector and each of the reference vectors from each topic was computed, according to Eq. (1). If a document scored above 0.25 when compared to a reference vector, that document's number and score were stored, along with which topic it scored well against. After all documents in the database were compared to all reference vectors, the documents were sorted by topic and score, duplicate documents within topics were removed, and a ranked list of documents gauged similar to at least one reference document in each topic was created.

One serious problem on this task was that the language and style of the reference documents was frequently quite different than that of the documents in the database. The reference documents were in large part drawn from newswire stories that presented a page or so of text discussing a single topic in some detail. The database, in contrast, was weighted towards documents with very different style and content. These documents included Federal Register documents, which tend to be quite large and generally quite diverse in topic and diffuse in style, as well as quite a bit of data from newsgroups, in which language was also quite unlike that of the reference documents.

The newswire documents were particularly difficult for Acquaintance to deal with. An example of some fairly typical texts in the newsgroups are shown in Figure 1 (names and addresses in the body of the text have been removed). The TEXT SGML tags separate the different messages.

```
<TEXT>
How do you place a transparent tint over a bitmap image in Photoshop
please?
 * SLMR 2.1a *
</TEXT>
 <TEXT>

I'm currently using QuarkExpress 3.3 for the Mac.  Is there a way to disable
hyphenation in a textbox?
</TEXT>
<TEXT>
I have perl5 Alpha 9, and when I run santa, I get this:
syntax error at perl/get_host.pl line 29, near "return $host_name_cache{$host"
syntax error at perl/get_host.pl line 32, near "else"

Can anyone shine light on it. Shall I get different version of perl
would you say.  Yours dissapointed after the hype.

</TEXT>
<TEXT>
```

A number of people have had trouble getting the short paper I wrote
on motion extrapolation. A preprint of it is given in PostScript
format below.

```
------------cut here----------------
%!PS-Adobe-2.0
%%Creator: dvips 5.495 Copyright 1986, 1992 Radical Eye Software
%%Title: motionextrap.dvi
%%Pages: 13
%%PageOrder: Ascend
%%BoundingBox: 0 0 596 842
%%EndComments
%DVIPSCommandLine: dvips motionextrap
%DVIPSSource:  TeX output 1993.07.06:1618
%%BeginProcSet: tex.pro
%!
/TeXDict 250 dict def TeXDict begin /N{def}def /B{bind def}N /S{exch}N /X{S N}
B /TR{translate}N /isls false N /vsize 11 72 mul N /@rigin{isls{[0 -1 1 0 0 0]
concat}if 72 Resolution div 72 VResolution div neg scale isls{Resolution hsize
-72 div mul 0 TR}if Resolution VResolution vsize -72 div 1 add mul TR matrix
currentmatrix dup dup 4 get round 4 exch put dup dup 5 get round 5 exch put
setmatrix}N /@landscape{/isls true N}B /@manualfeed{statusdict /manualfeed
true put}B /@copies{/#copies X}B /FMat[1 0 0 -1 0 0]N /FBB[0 0 0 0]N /nn 0 N
/IE 0 N /ctr 0 N /df-tail{/nn 8 dict N nn begin /FontType 3 N /FontMatrix
fntrx N /FontBBox FBB N string /base X array /BitMaps X /BuildChar{
CharBuilder}N /Encoding IE N end dup{/foo setfont}2 array copy cvx N load 0 nn
put /ctr 0 N[}B /df{/sf 1 N /fntrx FMat N df-tail}B /dfs{div /sf X /fntrx[sf 0
0 sf neg 0 0]N df-tail}B /E{pop nn dup definefont setfont}B /ch-width{ch-data
dup length 5 sub get}B /ch-height{ch-data dup length 4 sub get}B /ch-xoff{128
ch-data dup length 3 sub get sub}B /ch-yoff{ch-data dup length 2 sub get 127
sub}B /ch-dx{ch-data dup length 1 sub get}B /ch-image{ch-data dup type
```

Figure 1. Examples of texts from data for routing task

One problem was that many of the documents were so short that it was difficult to create a good statistical profile of them. Furthermore, the very unusual formats of some documents, as shown by the last example above, helped muddle the statistics on some files of documents. Paradoxically, had most or all the documents been in say, PostScript format, the system would have been better able to group them on the basis of content, as the PostScript "background" would have been accounted for and removed by the statistic profile created by the centroid. In any case, the content and style of these messages was very different from the newswire documents that characterized most of the reference documents.

In a effort to lessen the problems caused by the very different styles of language used in the reference documents and the documents from the database, two centroid vectors were used instead of one. First, a reference centroid vector from all of the reference documents was created. Then, documents from the database were read in one file at a time, and a centroid vector for that set of documents was created to capture the commonality among them. When comparing a document vector to a reference vector, the appropriate centroid was subtracted from the corresponding vectors, as shown in Equation 2:

$$S_{mn} = \frac{\sum_{j=1}^{J} (x_{mj} - \mu_j)(y_{nj} - \nu_j)}{\left[\sum_{j=1}^{J} (x_{mj} - \mu_j)^2 \sum_{j=1}^{J} (y_{nj} - \nu_j)^2 \right]^{1/2}} = \cos\theta_{mn}, \quad m = 1,\ldots,M, \;\; n = 1,\ldots N \qquad (2)$$

where the vectors $x_m,\ m \in 1,\ldots,M$ are the M document vectors, the vectors $y_n,\ n \in 1,\ldots,N$ are the N reference vectors in a *J*-dimensional space, μ is the centroid vector for the current file of documents from the database, and ν is the centroid for the set of reference documents.

The performance of the Acquaintance system on the routing track was rather poor. In fact, it performed significantly worse in TREC-4 than it did on the same track in TREC-3. In terms of average precision, it scored above the median only three times out of fifty. The reason for this was that in TREC-4 there was a much greater degree of mismatch between the documents that were used as references and the documents that were in the database. Since Acquaintance is a purely statistical system, if the statistics of the reference documents are significantly different from the documents in the database, it cannot perform well. In a real-world situation, if performance were this poor, one would add samples of documents whose content and style more closely modeled those in the database to the set of reference documents. The reference documents that were used for this task would be used only as a first approximation, and a set of more useful reference documents would either supplement or replace the original references.

Ad Hoc

The ad hoc task simulates the activity of a user who submits queries to a static database. The database is made available for the participants to train on early in the year, while the topic descriptions are only made available for a short time before the results of searches based on those descriptions are to be submitted.

In previous years the topic descriptions for the ad hoc task were fairly detailed. The topics consisted of a paragraph or two describing the topic, along with guidance as to what was and was not considered relevant to that topic, as well as a list of what amounted to keywords that helped define the topic even further. This year, the topics were very terse; in fact, some were almost telegraphic. For instance, topic 202 read “Status of nuclear proliferation treaties -- violations and monitoring.” On the other hand, some were more wordy, but actually much less specific, such as topic 216, “What research is ongoing to reduce the effects of osteoporosis in existing patients as well as to prevent the disease occurring in those unaffected at this time.” Logically, this topic boils down to “research on osteoporosis;” all other terms are redundant or uninformative. These extremely short topic descriptions are not untypical of spontaneous user queries, but by themselves they are not long enough from which to generate very solid statistics.

Due to the very sparse nature of this year’s queries, query generation was performed manually for all the ad hoc-based tasks. Since Acquaintance is a statistically-based algorithm, some minimum amount of vocabulary pertaining to the topic must be available for the system to reliably select documents with similar statistical profiles from a database. A few (usually no more than 5 or 6) words or phrases were therefore manually added to the supplied query, using the general subject knowledge of the users (Marc Damashek and Steve Huffman). That process took only a minute or two for each query. In addition, some terms deemed uninformative were removed. As an example, topic 201 originally read "What procedures should be implemented to ensure that proper care is given to children placed under the au pair's responsibility." This was changed this to read "au pair, children, proper care, nanny, nannies, caregiving, au pair, caretaker."

At this point, the modified queries were run against the documents in the database, and the highest scoring documents were returned. Those documents were then scored against each other. The 50 or so documents that were most highly connected to the other documents in the set, as determined by the Parentage tool, were automatically selected. These documents were then used as the reference documents for the final phase of scoring. If a document from the database scored above 0.25 when compared to the reference vector, that document's number and score were stored. Finally the documents were sorted by score, duplicate documents in each topic were removed, and a ranked list of documents gauged similar to at least one reference document was created.

The results on the ad hoc task in TREC-4 were considerably better than those in TREC-3. In spite of the sparseness of the queries, Acquaintance performed moderately well, scoring above the median in average precision on 15 out of the 49 topics. It would seem that the technique of running a first pass through the data to choose good candidate documents, and then using the most highly connected of those as the final set of reference documents, was more effective than last year's strategy of just using the given topic as the reference document.

Interactive

The interactive task permits the user of a system to interact with that system in a more natural fashion than the ad hoc task. The user is not limited to submitting a single query and simply accepting what the system returns. Rather, the user can examine the system's response to a query, and use that information to choose relevant documents, and/or further refine the query. The queries for this task were the a subset of those used for the ad hoc task.

There were actually two possible tasks for participants in this track. The first was simply to retrieve relevant documents, as in the basic ad hoc task. The second task was to use the system to create a new query, and submit the documents retrieved based on that query. The Acquaintance algorithm performed the first of these two tasks.

For this task, a somewhat different method was attempted than that used by most participants. A tool was used that shows the user the entire universe of documents that might be related to the topic at hand, and permits the user to roam through that universe, examining and/or selecting whole clusters of topic-related documents at one time. This is in contrast to those systems in which the user examines some set of documents returned by a system for a query, and then refines or resubmits the query based on the content of that set of documents.

This was accomplished with the Parentage information visualization system created by Dr. Jonathan Cohen (Cohen, 1995). For each topic, the 1000 top-scoring documents were found using the same procedure as the basic ad hoc task. Those 1000 documents were then scored against each other using the Acquaintance algorithm. Finally the documents and scores were submitted to the Parentage system, which graphed the relationships among the documents in that set.

In addition to visually mapping how documents cluster together, and how documents and clusters of documents relate to each other, the Parentage tool can automatically label each cluster of documents with a set of terms which characterize those words and phrases which cause that cluster of documents both to stand out from the rest, and pull together the documents within the cluster. These terms are referred to as "highlights." Parentage does this by using a modified version of the Acquaintance algorithm, using n-gram statistics and a form of centroid subtraction. An example of this can be seen in figure 2. This figure shows a screen shot of a small part of the Parentage graph for the documents from topic 242.

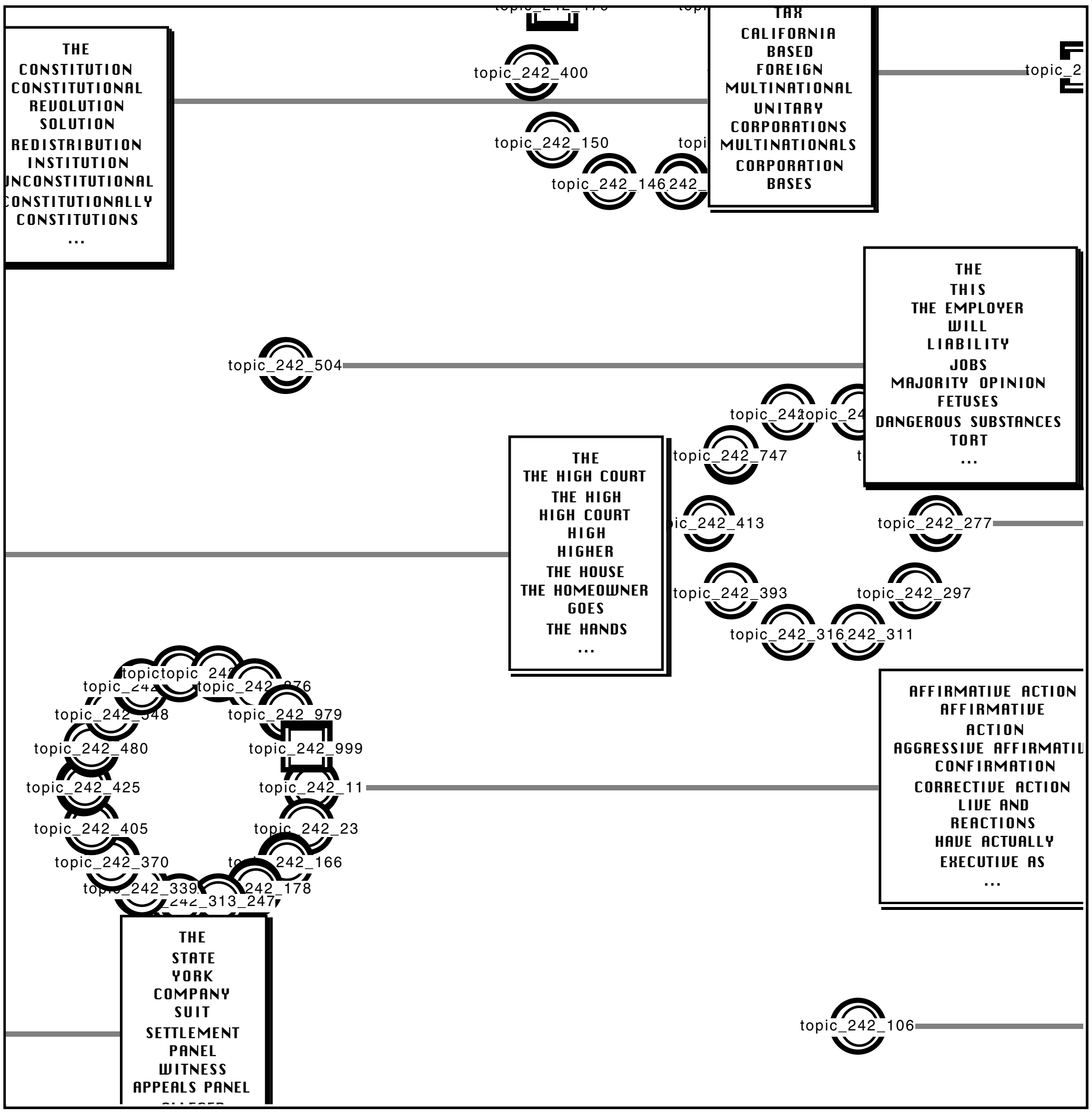

Figure 2. Portion of a Parentage display of a set of documents with automatically generated labels.

It can be seen in Figure 2 that each cluster of documents is shown with a list of highlights. Rather than needing to roam through the whole information space, the user can search for specific terms in either the highlights lists, or in the text of the documents themselves. This will put the user directly onto clusters of documents that may be of interest. Alternatively, if there is a good exemplar document for a topic, one can go directly to that document, and follow the paths of relationships leading from that.

In Figure 2, the user was looking for the term "affirmative action," since that was part of the subject of topic 242, which reads "How has affirmative action affected the construction industry?" By typing in the keyword "affirmative action," the user was moved directly to a cluster of documents dealing with that topic. The user at this point could continue searching for other clusters of documents, perhaps with other keywords, such as "construction." Once the potentially useful clusters of documents were isolated, the user can examine the documents individually, or merely select entire groups of documents. In cases where the highlights are suggestive but not diagnostic, the user can actually read all the documents in a cluster (all together or one at a time) in a window on the screen, and if the documents appear relevant, an entire cluster can be selected.

Another powerful feature of the Parentage tool is that the user can re-cluster sets of documents within their own context. For example, the cluster of documents on affirmative action can be moved to its own window, and the documents can be re-clustered and re-labeled based just on the text of the documents within that particular cluster. This effectively removes the common elements from the documents (in this case presumably terms dealing with affirmative action) leaving subtopics as the basis for clustering and labeling. Thus, if a user wanted a set of documents on a broad topic, such as affirmative action, the cluster so labeled in Figure 2 could be chosen with a fairly high degree of confidence that all the documents in it would be relevant to that topic. On the other hand, if the user wanted to find documents dealing with affirmative action in a certain context, Parentage could be used to examine subtopics within the overall cluster of documents on affirmative action.

It should be clear that whole clusters of related, relevant documents can be located and selected from the conceptual map in a remarkably short time. For any given topic, the users (again, Marc Damashek and Steve Huffman) needed to spend an average of less than ten minutes per topic finding the relevant documents; and for a few topics, they spent less than a minute locating and selecting the clusters of relevant documents. It was extremely easy to gather up the clusters based on the labels of probable content. In most cases, it freed the users from needing to read individual documents at all.

The use of an information mapping tool, in combination with a tool that measures document similarity (which need not be Acquaintance, but can be any system that characterizes the degree of similarity between two documents), is a very powerful method of exploring a database of documents. With such a system, one can understand the overall relationships among the set of documents. Unexpected relationships can be uncovered, and the centrality of certain documents is shown by the way that those documents draw together many disparate document clusters. The usefulness of such tools for dramatically enhancing both text retrieval and knowledge acquisition from a database is just beginning to be realized.

In terms of average precision, Acquaintance scored above the median in ten of twenty-five topics. That is not very impressive. However, informal results of the first task presented at the interactive panel session during the conference indicated that the performance of the Parentage and Acquaintance interactive system was very good, when other factors, such as the time to recover relevant documents, were taken into account.

Filtering

The object of the filtering task was to adjust a text retrieval system in such a way that it retrieved documents with high precision on one run, with high recall on another, and with a balance of precision and recall on a third. The data and queries used for these runs were the same as those used on the routing task.

The Acquaintance system attempted to achieve these three levels of performance by varying both the n-gram length and the threshold at which scores were reported. As n-gram width increases, the system obviously requires longer strings of text to be identical for them to be hashed to the same address in the document vector. By increasing n-gram length, and requiring a higher score threshold for defining documents as similar, the precision of the output should be increased.

For the high recall run, the n-gram length was set to five, and the score threshold was set at 0.25. This is actually close to typical parameters for using Acquaintance for topic based document retrieval. For the high precision run, the n-gram length was increased to seven, and the score threshold was increased to 0.40. This forced more and longer stretches of text to precisely match between the reference documents and the documents from the database to pass the threshold. For the balanced run, the n-gram length was kept at seven, but the score threshold was lowered to 0.30. This actually would result in a somewhat more stringent test for similarity than is normally used, but is still significantly less than the high precision run.

The results on this track were very poor, when compared to the other three systems that participated in the track. This is a reflection of the overall difficulty Acquaintance had with the mismatch in content and style between the reference documents for the routing task, and the documents in the routing database. It is not clear that better performance in comparison to the other systems could have been achieved by adjusting the parameters of the system given that fact.

Confusion track

This was a new track at TREC-4. Instigated in part because of interest by the defense community, this track was created to provide a vehicle for testing how text retrieval systems perform in the presence of garbled data. In the defense and intelligence worlds, data is often received in garbled form. Sometimes the garbling can be quite severe, and a system that cannot deal gracefully with degraded data is very limited in its usefulness.

The data for the corruption track consisted of the category B data, that is, a subset of the TREC data taken from Wall Street Journal and San Jose Mercury News articles. The data came in three forms, ungarbled, randomly garbled at ten percent, and randomly garbled at twenty percent. Random garbling meant that for any character in the text, there was a ten or twenty percent chance for that character to be changed, lost, or an additional random character inserted, with all garbles guaranteed to result in ASCII characters. Samples of the ten and twenty percent garbled text are shown in Figures 3 and 4. Only four systems participated in this track, and only Acquaintance and one other even attempted to process the data corrupted at the twenty percent level.

> To muc excitSement on top of too much cold iedZcation kmay have caused the
> pacpidPfjartbeatt t9haI forced Mansas Cidty linebacker DerricLC1 Thomas out of the
> Chiefs'plasyoff game SKturday.;   Thomas, a Pero OBowlhselection in all thre of
> hi6s yes in the NF, went out in the second qcartaer of the Chiefs' 10-6
> JiTctory over the LoYs Aygeuts RMidYrsshortly Zftey a sakH tht noced a fumQle.
> Sporwzs
> 3 RAPID EARTBEAT FORCES THOMAS TOEYAVE GAME
> OC. STAR IS EXPECTJD TO PLAY NEXoT WEEKEND
> Pro Football; AF8C Notebook

He kas takenito a hospital as a pre1Fution, although his heart rawte was
back to normal by thehtime he leftF the sadium. H7 remained ov9rnight for
observation.;   "The doctors indDMcaed Derrick may have taken too much cold
med7cation beforO te agakzmeJ," Chefs Prsident6Generna MVnager2Crl Peteron
sad. "That combined fwith t4he excitement of thegamef may haveS caused the
prblem.;   "We don'ti think it's anthin to be alarmezdabout.m;   Thomas is
expectdio be yable to play2next weekend.; SECOND-GESSING: RaiXer0s Coach Art
Shell refused to be secGnd-Ogussed about stBrting Todd Marinovichg at
quarterbsck ovr veteranpJay SchDoede.;   "ouF cwm doJit if you want," he
said, "but I'mhnot going to scond-gues mysJlUf.B;   SellF also bristled when
aUkd itf he considereT eplcing arinovich with Schroeder lateUin the game.;
"Mysthinking in the fourth 3quarter wisA fthaj we were here with the kid andwe
are going to finishGwi6th him," Shell sai.;   Schreder and YShell Ps5id that
Sch3oede5r, who sprained both ankles two9 weeks ago, was h4ealthy enoughCto
play.; COyTROVBRSIAL LAY: The second of Marinovich'svC fou interceptionsset
up the ony touchdwn oC the game.;   The score came on an 11-yard recepton
by the CXhiefs6' Fred Rones witpi 5 minutes, 7 slconds left i1 the second quartRr

Figure 3
Article 1 (SJMN91-06364024) from San Jose Mercury News at 10 percent garbling

Too much excitement n tp od toko mucx c3ld medicticon a ave causted9 te
rapid hertbeat thdat forced Kansas Cit l1inRblker Derric Thomaas sout4of te
hiefL' playsvf camee SaOtujaeI; gThomas, a Prao Bo6wl s7elction ino all thrQ oQ
Eis year in ahe NFL, wen ut in th sTelom quater oYf te Cifs' 0-Y
viNctory oer 6he LoswH Andgl7s Raiderssio2tlIatzer64 sack7that fotNed a fumble.
7 Y cports
e RAPID HEMATBAT FORCES THOMoAS TO LEAVh GAME
K.N. STAR 96S EXPkCTED TO PLAYNEXTWEEKND
Pro FootballX AF5Notebhok R
c jHe was ta9en to anhospital asAa pncaution, alth ough hims heart9b rMte was
beckqto nolmal y3the tie fe left thCsMtaNdium. Hesemaieed ovCrniht for
obsarvrtion.t 1TheV adokzctJors igndibyated rrickn vmy haypotaken tSoo7muce cYld
medicatfion be4fose the gafe," ChWibHfs PrxsSident-reneraklZManager Carl Pieterson
said. "FThaZt combined wzi1th4te excitement of the r5gabe8may havescauNd the
jroblem;   "WLk don'tMthnk 0it's anything ts be4alarmed wbotF"O; Q Thfjas i
expectedto be ale tog play next eckenud.; SEcND-GUESSNB: IRaidevrFs och At
Shell refusHd tocme seconyussed abouB oiarting TkoddMaUriovich 3at
Fquarterback over vetra4n8Jay Schroeoer. o "BYou 3can dPiL if you want,"mh
saidi "but I'm nt going tC second-YuUs3s myself.; A f S5ell also britled when
ased if 3he consfdered replaing6 Marinovichwith 7chrder xlate inYthevPaw;
"My ttnki5ngIin the fourth quarter was t9hat we were here8ith theGkid nd e
are goig D finibfhDwith hiY" Shell said.;Z vC Schkrceder aMd yShell said Jthjt
ScyhrodZr who spraibed both aneles two weeky ago, wxas wealtXhy e9ugJ to
lay.; CO5T ROVERz0AL PLAY The sepd oJMarinovioh'sLfur iterceptions set
0J tdhe only touchdown f the ga6me.;   The scoKeN4 cIme on an 11-yard receWtion
y the Chefs' Fced Jons with 5 m3inUutes, 7 seFconds luft inthe se od quarter

Figure 4
Article 1 (SJMN91-06364024) from San Jose Mercury News at 20 percent garbling

The uncorrupted and the ten percent corrupted text were processed in the same manner as the data in the basic ad hoc task. The n-gram length was five for both runs. The only change when processing the twenty percent garbled data was to change the n-gram length to four. This increased the chance that any particular n-gram would remain ungarbled.

Since Acquaintance is statistically based, some "noise" in the data should not cause the algorithm to fail catastrophically. In fact, Acquaintance performed very well on this task. It suffered minimal degradation in recall and precision between the uncorrupted and ten percent corrupted data; at ten percent garbling, Acquaintance scored above the median on thirty out of 49 topics. And while performance dropped again at the twenty percent corruption level, overall, the system still performed quite well. This indicated that the statistical nature of the algorithm let it degrade gracefully, and relatively slowly, as the data became more corrupt.

Spanish

The Spanish track was essentially the same as the English ad hoc track. Participants were given access to the Spanish database early, and then the queries were sent out shortly before the results were due back. The queries were, like the English queries, quite short. The database contained articles from El Norte, a Mexican newspaper.

The problem for Acquaintance here was the same as that when doing the ad hoc task in English. The topic descriptions were so short that they did not provide enough of a statistical profile to properly model the topics. A typical topic (number 32) read "Cual es la importancia de las Naciones Unidas (NU) para Mexico?" To overcome this, the topic descriptions were again manually expanded just from general subject knowledge of the users. Unfortunately, the users do not speak Spanish, and were not knowledgeable about Mexican affairs. Therefore, the "expansions" were very minimal, in fact usually consisting of removing clearly uninformative verbiage from the query rather than adding anything substantive to it. The rendering of the above query became "importancia de las Naciones Unidas (NU) para Mexico." Obviously, "query expansion" was of minimal use to Acquaintance in this track.

The Spanish ad hoc queries were processed in exactly the same manner as the English ad hoc queries. The results reflected the problems with the minimal queries; the performance of the system was quite poor. With fuller topic descriptions, or better manual query expansion, performance would have most likely have improved significantly, and in fact, should have been very comparable to the performance on the English ad hoc task.

Summary

The Acquaintance technique was developed to find documents that are similar to one another, or to a reference document, in a language independent and potentially garbled environment. For this to work acceptably as a topic spotting technique, it needs a modest amount of text in both the reference and the target documents that is relevant to that topic. In TREC-4, the queries in the ad-hoc based tasks were significantly sparser than in TREC-3, and this sparsity of text had an impact on the performance of the algorithm. Even so, the minimal manual augmentation of the topic descriptions, and the strategy of using the most highly connected documents from the first pass as reference documents helped improve the actual performance of the technique to the point that it outperformed last year's ad hoc results.

In the routing task, the documents against which the queries were compared were often either quite sparse and very different in style from the reference documents (the newsgroups), or quite diffuse (the federal register documents). This led to Acquaintance building very poor models from the reference documents of what was in the documents in the databases. The results in the routing-based tracks reflected this mismatch by the very poor performance of the system.

The system did perform quite well in the confusion track, which measures performance in an area where Acquaintance has a high degree of potential, namely, working with garbled data. Even at a relatively high degree of garbling, the system's performance degraded quite gracefully. This type of behavior is quite important to users of document retrieval and filtering systems in the defense and intelligence fields.

The other area where performance was rather good was in interactive document retrieval. This was achieved by the combination of Acquaintance with Parentage. The usefulness of information visualization for text retrieval, when combined with virtually any document retrieval engine, clearly has great potential.

References

[Cohen 1995] Jonathon Cohen: "Drawing Graphs to Convey Proximity: an Incremental Arrangement Method," submitted to ACM Transactions on Computer-Human Interaction.

[Damashek 1995] Marc Damashek: "Gauging Similarity via N-Grams: Language-Independent Categorization of Text," Science **246**, 843-848 (1995).

[Huffman 1995] Stephen Huffman, in preparation.