University of Surrey Participation in TREC 8:

Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)

Khurshid Ahmad
Lee Gillam
Lena Tostevin

AI Group
Department of Computing
School of EE, IT and Maths
University of Surrey
UK, GU2 5XH
K.Ahmad@surrey.ac.uk

Abstract

This paper describes the development of a prototype document retrieval system based on frequency calculations and corpora comparison techniques. The prototype, WILDER, generated simple frequency information based on which calculations of document relevance could be made. The prototype was built to allow the University of Surrey to debut in the U.S. Text Retrieval Competition (TREC).

User queries as specified by the TREC organisers were converted into simple word-frequency lists and compared against values for the entire corpus. These relative frequency values indicatively produced document relevance. The application of morphological and empirical heuristics enabled WILDER to produce the ranked frequency lists required.

Introduction

The *ad hoc task* of TREC8 investigates the performance of systems in ranking a static set of documents against novel topics (queries). For each topic, the top 1000 documents satisfying the topic are submitted. Recall and precision techniques are used on these rankings to determine the results of the competition overall.

We have used term identification and extraction techniques for identifying topics discussed in a given text. In this note we focus on the use of single word terms for identifying topics. The techniques are based on differences between general language texts, texts used in an everyday context, and special language texts. The special language texts are texts written, for instance, by scientists, engineers, business persons and hobbyists in their respective languages of physics, chemistry, engineering, business, and hobbies. English-speaking physicists will use the English rendering of terms of physics and use their knowledge of English language, which they share with other speakers of English. Similarly a Chinese speaking physicist writing in Chinese will use the Chinese rendering of terms plus their knowledge of Chinese which they share with other Chinese speakers. The special language texts can be distinguished from a collection of general language texts at different linguistic levels including lexical, morphological, syntactic and semantic. These differences can be measured quantitatively and qualitatively. Quantitative measures at the lexical level include frequency of usage of single and compound terms in special language texts and their equivalents in general language texts. Morphological differences can also be measured quantitatively by looking at the differences in the inflectional and derivational variants of terms; specialist texts comprise a larger number of plurals than used in general language; specialists use nominalised verbs more extensively than in general language.

The key difference at the lexical level, between specialist and general language texts, is in the distribution of the so-called open class words, typically nouns and adjectives, and the closed class words, typically determiners, conjunctions, prepositions and modal verbs. Consider the 100 million-word British

National Corpus (BNC) which 'was designed to characterise the state of contemporary British English in its various social and generic uses' (Aston and Burnard 1998); we will use the BNC as a general language corpus. The TREC 8 corpus in comparison to the BNC corpus can be regarded as specialist text corpus in that the former comprises financial and political news texts: about 30% of the text in TREC, measured by the number of documents, is derived from the London *Financial Times* and the other 45% is based on the *Federal Register* and *FBIS*. The potential general language component of TREC is based largely on the other 25% of the texts that are obtained from the *Los Angeles Times*. Tables 1a and 1b show the similarities and differences between the BNC and TREC-8 corpora in terms of the distribution of the 100 most frequently occurring tokens in the two. Note that the closed class words like determiners, prepositions and conjunctions have approximately the same distribution. The differences are in the number and appearance of the open class words. TREC-8 has 13 open class words, whereas the BNC can muster only 2. In the BNC, the first open class word *time* is the 79th most frequently word in the corpus, whereas in the TREC corpus the first open class word is *year* which is the 48th most used word in the corpus.

Table 1a. Distribution of 100 most frequent tokens in the British National Corpus (BNC comprises 4124 texts with over 100 Million tokens largely written and spoken during the 1970's and 1980's)

Tokens organised in order of frequency in batches of 10 at a time	Cumulative Relative Frequency	Number of Open Class Words
the,of,and,a,in,to,it,is,was,to	21.28%	0
i,for,you,he,be,with,on,that,by,at	6.66%	0
are,not,this,but,'s,they,his,from,had,she	4.35%	0
which,or,we,an,n't,'s,were,that,been,have	3.25%	0
their,has,would,what,will,there,if,can,all,her	2.42%	0
as,who,have,do,that,one,said,them,some,could	1.90%	0
him,into,its,then,two,when,up,**time**,my,out	1.57%	1
so,did,about,your,now,me,no,more,other,just	1.37%	0
these,also,**people**,any,first,only,new,may,very,should	1.18%	1
as,like,her,than,as,how,well,way,our,as	1.02%	0
Total Text (100106029 tokens)	45.01%	2

Table 1b. Distribution of 100 most frequent tokens in the TREC-8 Corpus (The corpus comprises 528155 texts with over 600 Million tokens largely written the 1990's)

Tokens organised in order of frequency in batches of 10 at a time	Cumulative Relative Frequency	Number of Open Class Words
the,of,to,and,in,a,for,that,is,s	22.36%	0
on,with,by,be,it,as,at,was,are,from	5.47%	0
this,said,will,has,not,have,he,an,or,which	3.76%	0
but,its,i,they,we,his,would,**year**,been,their	2.40%	1
were,who,one,had,more,**mr**,all,1,new,**per**	1.88%	2
there,no,also,about,up,than,other,if,hyph,**government**	1.58%	1
two,**cent**,may,out,when,after,2,last,**state**,0	1.34%	2
first,**pounds**,**people**,only,can,you,**time**,some,over,**company**	1.21%	4
into,such,**market**,should,any,under,**years**,so,us,these	1.05%	2
what,t,3,because,**ft**,94,do,could,most,now	0.93%	1
Total Text (255637339 tokens)	41.98%	13

Table 1c shows the distribution of the open and closed class words in the various sub-corpora of the TREC –8 corpus. It appears that the *Federal Register* has the largest number of open class words amongs its first 100 words, followed by *FBIS*, and the *FT*. *LA Times* behaves differently in that it has only a 1/3rd of the open class words amongst its 100 most frequent words when compared to a similar number in the *Federal Register*. Recall that the BNC has only 2 open class words amongst the 100 most frequent words: A simple χ -square test will show that these subcorpora are different from the BNC on the basis of the frequency of open class words amongst the 100 most frequent words.

Table 1c. Distribution of open and closed class words in the TREC subcorpora

Group	*Federal Register*	*FBIS*	*Finacial Times*	*LosAngeles Times*	Row Total
(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
Open Class	**33**	**26**	**21**	**11**	**91 (22.7%)**
Closed Class	**67**	**74**	**79**	**89**	**309 (77.3%)**
Column Total	**100**	**100**	**100**	**100**	**400**

It has been argued elsewhere that there are substantive differences at the morphological level in the use of keywords and certain verbs in the more formal literature of science and technology when compared to general language texts (see, for instance, Biber, Conrad and Reppen 1998). In the *FBIS* subcorpus the inflected forms *countries, elections,* and *relations*, and the derived forms *European, Russian,* and *Spanish* are respectively more frequent than *country*, *election*, *relation*, *Europe, Russia and Spain*. Similarly in the FT subcorpus *pounds*, *dollars* and *shares* are more frequent than their singular forms; and, there is little difference in the frequency of *company*/*companies* and *share/shares*. (In the BNC we note that *shares* are more frequent than *share*, but the term *dollar* is used 4 times more than the plural form). The *LA Times* subcorpus, however, does not have the same characteristics in that not only it has only 11 open class words amongst the 100 most frequent words, it has no plurals or nominalised verbs either amongst the 100 most frequently used words.

The lexical and morphological differences can help in filtering closed class words from special language texts and also certain commonly used open class words. This filtering process, should in principle, will result in a list of words that may be more closely related to the topic or theme of the paper. Some of the open class words or terms are usually *carriers* of meaning in that such words are used generally as a part of a complex phrase; for instance, the term *virus*, is used frequently in virology texts but occurs mostly as a part of a compound like *African Green Monkey virus* or *AIDS virus*. The meaning associated with the stem *virus* is related to the context of its usage in specialist texts. Similarly, the term *dollar* does not convey much information in international finance texts unless the context is examined, for example, whether the author of a given text was discussing *US $*, *Australian $* or *dollar-denominated bonds.* The following example illustrates the point made above. This is especially true if the specialist lexical item has entered general language vocabulary

We have carried out an experiment in which we removed the first 100 and then first 2000 most frequently occurring words in the BNC from the frequency lists compiled from the *FT, FBIS*, *LA Times* and the *FR* subcorpora. Tables 1d and 1e show the filtered wordlists from the *FT* subcorpus after the 100 and 2000 words from the BNC were excluded from the *FT* lists.

Table 1d. The residual, frequency ordered wordlist for the FT subcorpus after the first 100 most frequently words (occuring in the BNC) were removed.

mr,per,cent,pounds,year,ft,company,market,us,last	0.033063322
dollars,government,over,group,uk,yesterday,0,after,1,companies	0.016034449
bank,years,most,business,says,such,international,shares,world,2	0.010966193
however,news,tax,european,between,94,week,industry,93,share	0.009264547
three,interest,next,against,sales,profits,92,investment,while,london	0.008306062

Table 1e. The residual, frequency ordered wordlist for the FT subcorpus after the first 2000 most frequently words (occuring in the BNC) were removed.

cent,ft,dollars,0,94,93,92,markets,amp,investors	0.015253328
trading,cut,chief,index,finance,earnings,net,fall,turnover,japanese	0.004587571
according,losses,pre,announced,inflation,increased,non,debt,least,operating	0.0035461911
recovery,dividend,average,bond,spending,china,talks,recession,biggest,co	0.002904808
equity,currency,stake,official,analysts,trust,shareholders,assets,businesses,securities	0.002577078

Table 1d shows some of the keywords that form the basis of the English variant of the special language of finance and commerce. *FT* it appears focuses on *dollars*, *pounds*, *shares* and *industry*. Table 1e shows that when we remove the first 2000 most frequent words from *FT's* wordlist we are dealing with more specific issues like *markets*, *investors*, *earning* and *losses*.

Weirdness of special language texts

The differences in the distribution of certain lexical items, and their variants, in special and general language texts can be quantified in terms of the relative frequencies of a specialist text (corpus) and a general language text corpus. We call this ratio an index of *weirdness* of a specialist text. This weirdness is used by an accentuated, and perhaps an eccentric, choice of lexical items measured in terms of their frequency of occurrence. Most weird words in a text will tend to represent it more closely than those that are not as weird. If the ratio is unity, then the lexical item has the same frequency in both general and special language; if the ratio is greater than unity then the item is used more frequently in specialist text then is the case for general language and vice versa. (The anthropologist Bronsilaw Malinowski used the term *weird* to describe the language of *shamans* of South Sea Islands because they were using lots of names of spirits and objects).

It can be argued that comparison of the frequency distribution of items in special-language and general-language texts can identify signatures of a specialism. This technique has the advantage of being language-independent once the general-language corpus - or even a frequency list - has been obtained. 'Closed-class' words will tend to have ratios of around 1:1 in this comparison whereas terms or term carriers - content words rather than form words - will have a much higher ratio since their frequency in general-language texts will be low or potentially zero.

$$\boxed{Weirdness = \frac{w_s / t_s}{w_g / t_g}}$$

Where: w_s = frequency of word in specialist language corpus
w_g = frequency of word in general language corpus
t_s = total count of words in specialist language corpus
t_g = total count of words in general language corpus

Consider the weirdness coefficients of some of the most frequent terms used in the *TREC-8* corpus; we have used BNC relative frequencies to compute the ratio.

	Freq (BNC)	Rel Freq (BNC)	Freq (TREC-8)	Rel Freq (TREC-8)	Weirdness
Dollar	2023	2.02086E-05	30450	0.000119114	5.894233822
Dollars	1677	1.67522E-05	182147	0.000712521	42.53289129
Government	62163	0.000620972	383115	0.001498666	2.413421274
Governments	4731	4.72599E-05	26413	0.000103322	2.186254565
Islam	523	5.22446E-06	4108	1.60696E-05	3.075846739
Islamic	1290	1.28863E-05	19410	7.59279E-05	5.892122549

Market	23719	0.000236939	278277	0.001088562	4.594273903
Markets	3895	3.89087E-05	80749	0.000315873	8.118310106

Recall the differences between the TREC corpus and the BNC. The BNC is a weighted corpus containing much diversity of general language texts so that no specific topic or domain has dominance. In the TREC corpus, governmental, financial and personal information are highly frequent, evident in the number of nouns occurring in the top 10 percentiles above. These 100 tokens make up 45% and 42% of the entire collection of the texts, representing 45,057,724 and 107,324,924 tokens respectively.

An immediate consequence of this fact is that analysis of the TREC corpus is considerably varied in contrast to that of the British National Corpus, biased towards these nouns. This information needs to be factored out of any contrastive analysis within the data.

Method

In order to compute the relevance of a given text to a query posed in TREC-8, the following steps shown in figure 7.1 were taken:

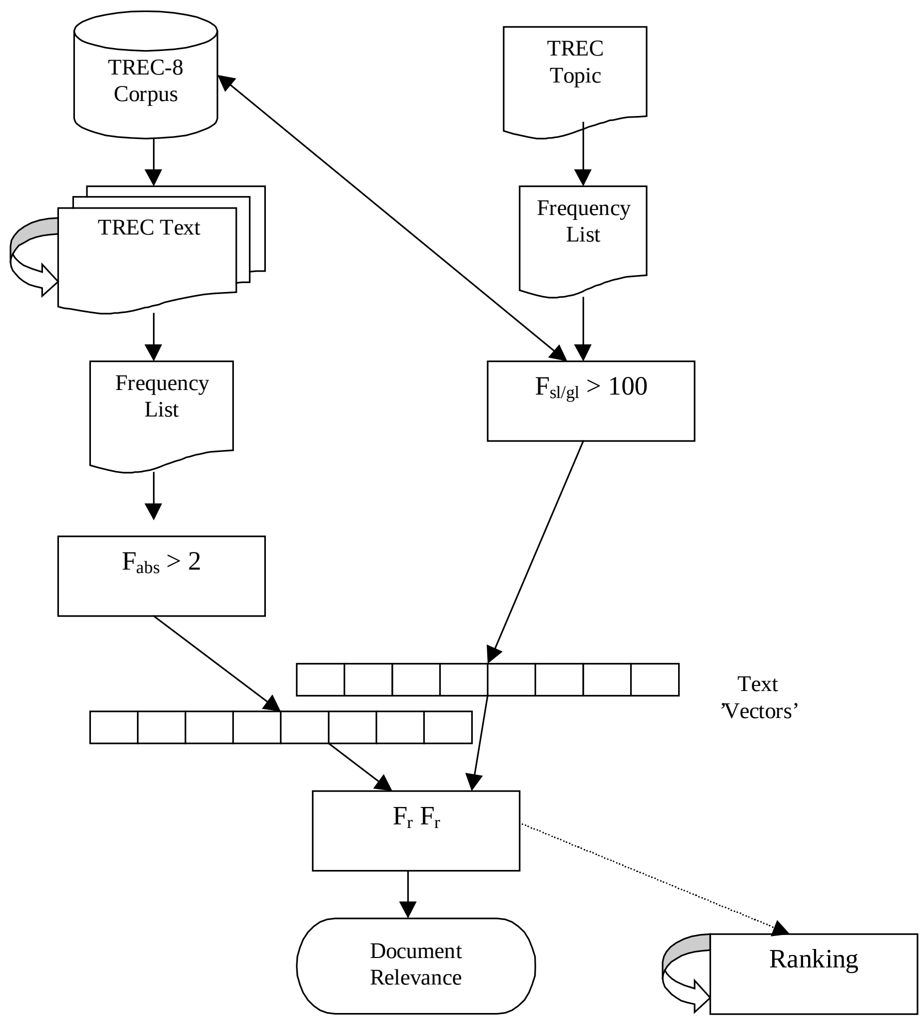


Figure 1: Steps to relevance

The resulting 'Vectors' - text and topic were then compared using the following correlation:

$$\sum fr_{sl}\, fr_{gl} - \sum fr_{sl}^{m}$$

Relevance was computed over the TREC-8 corpus for each topic and the texts were ranked. The 1000 most relevant texts were selected and submitted.

Consider Topic 444 in TREC-8:

```
<top>

<num> Number: 444
<title> supercritical fluids

<desc> Description:
What are the potential uses for supercritical fluids
as an environmental protection measure?

<narr> Narrative:
To be relevant, a document must indicate that the
fluid involved is achieved by a process of pressurization
producing the supercritical fluid.

</top>
```

After removing words with Weirdness <= 100 we obtain the following weirdness-ordered wordlist:

WORD	FREQ	ABS FREQ	WEIRDNESS
achieved	1	0.0227	276.0000
document	1	0.0227	182.0000
fluid	4	0.0910	4780.0000
indicate	1	0.0227	590.0000
involved	1	0.0227	135.0000
measure	1	0.0227	249.0000
Potential	1	0.0227	133.0000
pressurization	1	0.0227	50500.0000
producing	1	0.0227	469.0000
protection	1	0.0227	122.0000
relevant	1	0.0227	370.0000
supercritical	3	0.0682	236000.0000
uses	1	0.0227	413.0000

TREC-8 determined the following texts to be relevant to this topic:

```
FBIS4-20472                        FBIS4-44730
FBIS4-44741                        FBIS4-44747
FBIS4-44913                        FBIS4-45803
FBIS4-66450                        FR940128-2-00102
FR940318-0-00170                   FR940318-0-00172
FR940318-0-00173                   FR940318-0-00213
FR940607-0-00051                   FR940721-2-00028
FT932-7115                         FT933-14063
FT943-4354
```

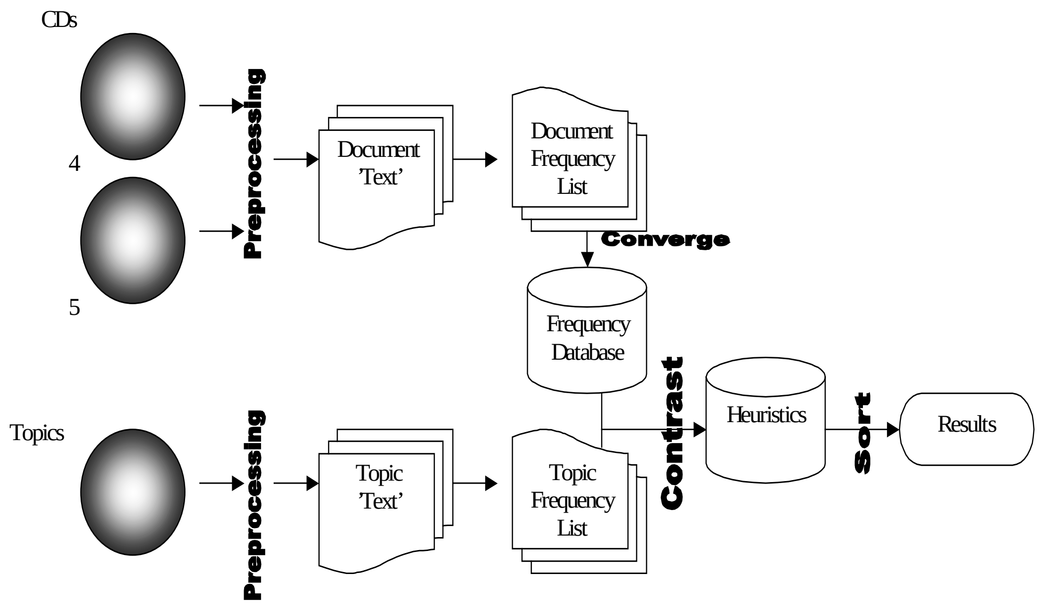
Of these texts, in the first 10 that we selected, we now know that the texts marked in bold were relevant to this query:

```
444 Q0       FBIS4-20472 0  0.000471 surfahi2
444 Q0       FBIS4-44913 1 -0.051765 surfahi2
444 Q0       FBIS4-44747 2 -0.207757 surfahi2
444 Q0       FBIS4-45803 3 -0.207972 surfahi2
444 Q0  FR940318-0-00173 4 -0.208627 surfahi2
444 Q0       FBIS3-41666 5 -0.209205 surfahi2
444 Q0  FR941206-1-00134 6 -0.209513 surfahi2
444 Q0       FBIS3-40501 7 -0.209580 surfahi2
444 Q0  FR940812-2-00056 8 -0.209600 surfahi2
444 Q0       FBIS3-40450 9 -0.209660 surfahi2
```

The WILDER program

In order to participate in this task, a prototype system, WILDER was developed. The system was built from a combination of existing Java, Perl and C code, and Unix shell scripting and associated utilities to achieve significant performance and ease of development.

This architecture for WILDER is shown below, which allows for a number of modular elements which can be developed in parallel and allows for a number of contrast algorithms to be switched in and out of the model in order to evaluate specific hypotheses.



Run stats

All processing was done within the Sun Solaris system. Building the original comparison resources took approximately 4 actual days on a single Sparc Ultra 1 - 140. Subsequently, each query took approximately 8 hours to satisfy the query from the raw results. There are many available optimisations to the algorithm used.

Future Direction

We have argued that fully automated extraction system can be created using simple contrastive frequency techniques for Information Retrieval in order to identify the topic of specific texts. The relative length of each text - at an average of 3.6K - is indicative of a lack of intra-text synonymy or term variants as would be true of lengthy narrative reports.

Treatment of simple morphology, acronyms, proper names and abbreviation needs further consideration within this particular arena, as does the application of techniques such as LSI and raw synonymy. Potentially, varying the values chosen for the application of the heuristics may make improvements to this simple methodology. Our goal was to build a system capable of handling such volumes of text within workable time. It is now our goal, based upon the results we have achieved, to improve and optimize this system using we have learnt through participation in this competition.

References

Aston, Guy and Burnard, Lou. (1998). *The BNC Handbook: Exploring the British National Corpus*. Edinburgh: Edinburgh University Press.

Biber, Douglas, Conrad, Susan and Reppen, Randi. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Salton, G and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill