Assessed Relevance and Stylistic Variation

Jussi Karlgren
Courant Institute of Mathematical Sciences
Department of Computer Science
New York University
715 Broadway # 704
New York, NY 10014
karlgren@cs.nyu.edu

February 1996

Abstract

Texts exhibit considerable stylistic variation. This paper reports an experiment where a large corpus of documents is analyzed using various simple stylistic metrics. A subset of the corpus has been previously assessed to be relevant for answering given information retrieval queries. The experiment shows that this subset differs significantly from the rest of the corpus in terms of the stylistic metrics studied.

1 Introduction

Texts vary not only by topic. Indeed, stylistic variation between texts of the same topic is often at least as noticeable as the variation between texts of different topic but same genre or variety.

This experiment compares simple measurements, indicative of stylistic variation, on a corpus of documents, with measurements made on a subset of documents that have previously been judged relevant for answering queries from a given set.

The Text REtrieval Conference (TREC), organized in the form of a competition by ARPA and NIST, gives participating organizations access to a large corpus of texts and a set of queries that are to be used for retrieving texts from the corpus. Of the texts that are retrieved by the participating information retrieval systems, a certain number are read by a number of human judges, and assessed as relevant or not relevant. Thus, given a query, the corpus is partitioned in three parts: relevant texts, not relevant texts, and not assessed texts (Harman, 1995).

For this experiment a corpus of ninety thousand documents was randomly selected from the TREC corpus[1]. A corpus of thirty thousand documents was similarly selected for testing purposes. The breakdown per source category can be seen in table 1.

Initially, the documents were analyzed for simple word and sentence statistics, such as are used in readability analyses (Klare, 1963), a method which has been used for investigating style and genre variation in the past (Biber, 1988, 1989; Karlgren and Cutting, 1994). Subsequently the texts were analyzed for subtopic structure (Hearst and Plaunt, 1993), and for syntactic complexity, using a robust parser developed for information retrieval applications (Strzalkowski, 1994).

2 Results

The results are positive. The hypothesis of the experiment was that relevant texts in this sort of homogenized scenario would differ systematically from texts which are not relevant.

[1]The material was taken from TREC Disk 2, with the addition of San Jose Mercury News from TREC Disk 3.

Source	Number	Relevant	Misses	Not Judged
Associated Press Newswire	23766	374	2522	20870
San Jose Mercury News Articles	25075	267	3593	21215
Wall Street Journal Articles	22434	230	3948	18256
Ziff-Davis Computer Select Articles	17183	42	880	16261
Total	88458	913	10943	76602

Table 1: Training corpus composition

This turned out to be the case, and for most metrics tested, the difference was striking. But in addition, we find that relevant texts and non-relevant texts taken together – i.e. texts highly ranked by systems participating in the TREC evaluation – differ from the rest of the corpus in a systematic manner. The difference between relevant and non-relevant texts is much smaller than the difference between either of them and the non-judged portion of the corpus, but still significant even by univariate criteria in several of the metrics inspected. As a significance test we use the Mann Whitney U rank sum test.

In summary, the results of this experiment show that retrieved highly ranked texts – both relevant and non-relevant – are longer, with a more complex sentence structure, and with a larger number of subtopics, than the rest of the corpus. Relevant documents differ from non-relevant in a more convoluted way. Long relevant documents seem to be simpler – as regards sentence and subtopic structure – than long non-relevant documents; short relevant documents, on the contrary, seem to be more complex.

2.1 Simple statistics: Sentence Length and Word Statistics

A simple word count reveals that relevant texts on the average are longer than other texts – which also has been observed, pointed out, and utilized by the very successful Cornell research group at the latest TREC conference (Buckley et al., 1995). This is at least partly due to the fact that longer texts range over several topics, and thus there is a chance that a long text will touch a relevant topic. In this experiment, we find that not only are relevant documents longer, but all documents retrieved by systems, even those assessed by human judges as irrelevant, also are longer than the average document. Not only will longer texts touch relevant topics – but apparently they may well touch irrelevant but confusingly similar topics. On closer inspection, this is not entirely surprising. The non-retrieved portion of the corpus turns out to contain large numbers of very short items, and large numbers of tables and statistics, both short and long, which the retrieval systems have not proffered to the assessors for consideration.

Relevant texts also have longer sentences and longer words. Word statistics – word length, long word counts, type/token ratios – as a measure of terminological complexity have often been paired with sentence length to produce readability scores (Klare, 1963) or genre discrimination metrics (Karlgren and Cutting, 1994). We will return to discuss syntactic complexity in a separate section below, but note that in order to control for the fact that a large number of non-assessed texts were very short, the experiment was run again, this time on texts in different length categories: under one hundred words, between one and two hundred, between two and five hundred, between five hundred and one thousand, and over one thousand words in length. The differences between categories as regards sentence length persisted – most probably attributable to tables and stock market listings and other not very textual data – as did the difference in word length. Type/token distinctions did not, as might be expected. The difference between relevant and non-relevant texts is highly significant even on an univariate test. Table 2 contains a summary of results. The differences between relevant and non-relevant are significant in a Mann Whitney test on a 95% confidence level when marked with an asterisk in the table.

Category	Number	Word count	Words per sentence	Word length	Type-token ratio
All	31823	445	15.0	4.94	0.5776
Relevant	1327	*650	*17.2	*5.04	*0.5223
Misses	6063	*612	*15.8	*5.01	*0.5434
Not judged	24433	392	14.7	4.93	0.5891

Table 2: Sentence length averages

Category	Number	Tiles
All	32193	2.2
Relevant	1337	*3.3
Misses	6138	*3.2
Not assessed	24718	1.93

Table 3: Average number of tiles

2.2 Subtopic structure

Texts that are relevant are longer – and may be so for several reasons. One reason, as discussed above, is that they may range over several subtopics. We will here test this assumption, by comparing the relevant, non-relevant, and not judged portions of the corpus using a metric for computing subtopic shift. The text tiling algorithm (Hearst and Plaunt, 1993) partitions a text into probable subtopic chunks based on changes in word occurrence statistics. While the results the algorithm produces may be less than absolute – subtopic is not an objectively evaluable concept, and there are typically several ways of segmenting a text into subtopical passages – it does give an indication of textual type differences and terminological drift in texts. We find a clear difference between either relevant or non-relevant texts on the one hand, and the rest of the corpus on the other as shown in table 3. The differences between relevant and non-relevant are significant in a Mann Whitney test on a 95% confidence level when marked with an asterisk in the table.

Now, document length will affect the subtopic structure. If we partition the corpus in different length segments to see how, we find something very curious: relevant documents tend to have slightly more subtopics than irrelevant ones, if the analysis is restricted to short documents. For longer documents, the distinction is the opposite: long relevant documents tend to have fewer subtopics than long irrelevant ones. See table 4.

Documents with 200-500 words		
All	1946	1.31
Relevant	372	1.33*
Misses	1574	1.31*
Documents with 500-1000 words		
All	2702	3.4
Relevant	602	3.4
Misses	2100	3.4
Documents over a thousand words		
All	1245	8.6
Relevant	205	8.0
Misses	1040	8.7

Table 4: Tile Counts For Documents Of Different Lengths

2.3 Syntactic complexity

Syntactic complexity is a dimension which exhibits considerable variation between genres (Menshikov, 1974; Losee, forthcoming). Indeed, most stylistic measures heretofore have been attempts to find a shortcut to measure syntactic complexity; sentence length, like used above is one such method, although arguably a blunt one – what syntactic constructions are complex in themselves, and when they are evidence of complexity in an already complex subject matter is a matter of contention and psycholinguistic study (cf. Dawkins, 1974).

As a simple approximation of clause complexity, we will look at the average depth of output trees from a robust parser built for information retrieval purposes (Strzalkowski, 1994). In addition, the parser was set to skip parsing after a timeout threshold, and when it does so, it notes it has done so in the parse tree. These skip marks were counted – again, as an indication of clausal complexity. We find below, in table 5, a clear distinction

Category	number	depth	skips
All	32193	235	8.30
Relevant	1337	323*	12.4*
Non-relevant	6138	312*	11.9*
Not assessed	24718	211	7.17

Table 5: Trees and Skips

Category	number	depth	skips
Documents under a hundred words			
All	597	76.8	1.22
Relevant	34	79.6	1.79*
Misses	563	76.7	1.19*
Documents with 100-200 words			
All	900	109	2.96
Relevant	114	113*	3.21
Misses	786	109*	2.92
Documents with 200-500 words			
All	1946	191	6.59
Relevant	372	194	6.77
Misses	1574	190	6.55
Documents with 500-1000 words			
All	2702	357	13.5
Relevant	602	350*	13.1*
Misses	2100	359*	13.6*
Documents over a thousand words			
All	1245	672	28.9
Relevant	205	633*	27.3*
Misses	1040	680*	29.2*

Table 6: Trees and Skips For Documents of Different Lengths

between the various categories of document. Relevant documents have, on average, deeper parse trees and more skips. The difference between relevant and non-relevant is significant in a Mann Whitney test on a 95% confidence level when marked with an asterisk in the table.

Again, inspecting documents in classes of different length we find, as in the case with the subtopic analysis, that long relevant and short relevant documents are different from irrelevant ones in different ways. Table 6 shows how short relevant documents have more misses and deeper parse trees than short irrelevant ones; long relevant documents have fewer misses and shorter parse trees than irrelevant ones.

3 Conclusions

Texts differ in style. In this case, not very surprisingly, the retrieved texts differed from the main corpus along several metrics. What is more interesting, and may prove useful in information retrieval application, is utilizing this type of measure in distinguishing relevant texts from less relevant ones. This will entail analyzing the tasks and expectations of users; this experiment shows that for a certain set of users and for a certain scenario a clear bias towards a certain types of text can be found.

The differences between relevant and non-relevant texts found should not be taken as general results: while useful in a TREC context, as shown by the results from Cornell, they are clearly an effect of the task, corpus, and assessors. These results should be taken as a starting point in investigating how situations affect measures of stylistic variation.

References

Douglas Biber. 1988. *Variation across speech and writing.* Cambridge University Press.

Douglas Biber. 1989. "A typology of English texts", *Linguistics, 27*:3-43.

Chris Buckley, Amit Singhal, Mandar Mitra, Gerard Salton. 1995. New Retrieval Approches Using SMART: TREC 4. In Proceedings of TREC-4.

John Dawkins. 1975. *Syntax and Readability.* Newark, Delaware: International Reading Association.

Donna Harman. 1995. Overview of the Third Text REtrieval Conference (TREC-3). In Proceedings of TREC-4.

Marti Hearst and Christian Plaunt. 1993. "Subtopic Structuring for Full-length Document Access". Proceedings of the 16th ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh. New York: ACM.

Jussi Karlgren and Douglass Cutting. 1994. "Recognizing Text Genres with Simple Metrics Using Discriminant Analysis", *Proceedings of COLING 94*, Kyoto. (In the Computation and Language E-Print Archive: cmp-lg/9410008).

George R. Klare 1963. *The Measurement of Readability*,Iowa Univ press.

Robert M. Losee. forthcoming. Text Windows and Phrases Differing by Discipline, Location in Document, and Syntactic Structure. *Information Processing and Management.* (In the Computation and Language E-Print Archive: cmp-lg/9602003).

I. I. Menshikov. 1974. "K voprosu o zhanrovo-stilevoy obuslovlennosti sintaksicheskoy struktury frazy". In *Voprosu statisticheskoy stilistiki.* Golovin et al. (eds.)

1974. Kiev: Naukova dumka; Akademia Nauk Ukrainskoy SSR.

Tomek Strzalkowski. 1994 "Robust Text Processing in Automated Information Retrieval". Proceedings of the Fourth Conference on Applied Natural Language Processing in Stuttgart. ACL.

Donna Harman (ed.). Forthcoming. Proceedings from the Fourth Text REtrieval Conference (TREC-4). Gaithersburg, Maryland: NIST.