EXPERIMENTS IN MULTI-LINGUAL INFORMATION RETRIEVAL

G. Salton

TR 72-154

December 1972

Computer Science Dept.
Cornell University
Ithaca, N.Y. 14850

Experiments in Multi-lingual Information Retrieval

G. Salton

Abstract

A comparison was made of the performance in an automatic information retrieval environment of user queries and document abstracts available in natural language form in both English and French. The results obtained indicate that the automatic indexing and retrieval techniques actually used appear equally effective in handling the query and document texts in both languages.

1. Background

In an earlier study, the automatic text analysis and retrieval techniques incorporated into the SMART system were used to process documents and queries in both English and German. A multi-lingual thesaurus was utilized, containing entries in both English and German, and it was found that the manual translation of user queries from one language to another (from English into German) did not affect their retrieval performance. Specifically, when the original English queries and their translated German equivalents were processed against an English document collection the same retrieval performance was obtained for both sets of queries. This was also true when the two query sets were processed against a collection of German documents. [1]

+ Dept. of Computer Science, Cornell University, Ithaca, N.Y. 14850
This study was supported in part by the National Library of Medicine (NIH) under grant LM 00704.

Recently, a similar analysis was carried out with a Russian document retrieval system, known as the "Empty-Nonempty 2" system. In that case, English as well as Russian versions of a "descriptoral dictionary" (thesaurus) were constructed, and the performance of an identical set of queries was evaluated for both English as well as Russian language versions. Again, the retrieval results indicated that the performance was comparable for the queries in both languages. [2]

While these earlier tests showed that multi-lingual analysis tools could be constructed which would operate equally well for several natural languages, the retrieval results obtained for the document collections in the various languages could not be compared directly, because the documents used differed from one collection to another — the queries alone being identical for the respective pairs of languages.

In the present test, this earlier flaw is removed: the collections of document abstracts as well as the queries are exactly the same in both languages. This makes possible a direct comparison of retrieval runs for both the English as well as the French collections. In the present case also, French replaces the earlier German as the foreign language. Finally, a new automatic term weighting process not available earlier is used in the analysis of the French and English vocabularies to assign high weights to those terms which are best able to distinguish the documents from each other (the best document discriminators).

2. The Experiment

The experimental environment is summarized by the data of Table 1. An identical set of 52 document abstracts in the area of documentation was made available in both French and English versions by the Euratom Common Research Center in Ispra (Italy). This collection was used with 16 user queries also available in French and English. A thesaurus, or synonym dictionary, was constructed manually for each language, to group sets of related, or synonymous, words into individual thesaurus classes, the classes being numbered in such a way that corresponding word groups in both languages were assigned the same class identifiers.* Thus terms such as "language" and "syntax" are grouped into thesaurus class 255 in both thesaurus versions. A thesaurus excerpt is reproduced in Fig. 1.

It may be seen from Table 1 that the French vocabulary is considerably more diversified, since 1340 distinct French word types correspond to only 1197 in English. This is reflected in the thesaurus where 323 thesaurus classes are used for French, compared with only 287 for the English version.

The following main automatic analysis techniques were used to reduce the original natural language versions of both document abstracts and query statements into analyzed concept vector form: [3]

---

*The English thesaurus was generated by Barbara Galaska at Cornell University, whereas the French version was handled by Viviane Guiette-Limbourg of the Centre National de Documentation Scientifique et Technique in Brussels (Belgium). The assistance of both individuals is gratefully acknowledged.

a) the word form process eliminates certain common function words by using a stop list, and cuts off final "s" endings so as to reduce to the same form both singular and plural versions of the same word; a weight is automatically assigned to each word form based on the frequency of occurrence of the word in each document or query statement;

b) each word form can be replaced following a search process by the corresponding thesaurus class numbers; at the same time, weights may be assigned to the thesaurus classes, derived from the weights of the original component word forms;

c) a discrimination value can be computed for each word form based on the amount of discrimination supplied by each word when assigned to the documents of a collection; specifically, for a given document collection, it is possible to compute the average inter-document similarity (the average matching coefficient between pairs of document vectors) first without having term x present in the document vectors, and later with term x assigned; the discrimination value for term x is then a function of the differences in inter-document similarities: if the similarity between documents decreases when term x is assigned, x is rated as a good discriminator with a positive discrimination value; contrariwise, if the inter-document similarity increases with term x, x is a bad term with a negative discrimination value. [3]

Given the document and query vectors constructed in accordance with one or another of the methods previously described, it is possible to compute a correlation (similarity) measure between query and document vectors, followed by the retrieval of those documents which exhibit sufficiently high correlation coefficients. Furthermore, if relevance judgments are available which assess the relevance of each document with

respect to each query, the normal recall-precision output can be produced to evaluate the effectiveness of the retrieval output in terms of relevant items properly retrieved, and nonrelevant material correctly rejected.

In the present experiments, relevance assessments were obtained from the query authors, and retrieval runs were performed for the standard word form vectors; a modified word form match in which the weight of each word is multiplied by the corresponding discrimination value (Word Form - D.V. Weight), the standard thesaurus vectors, the thesaurus method with class weights multiplied by the corresponding class discrimination values (Thesaurus - D.V. Weight), and, finally, for the thesaurus vectors with negative discriminators — classes with negative discrimination values — eliminated (Thesaurus - Negative Discriminators).

3. Evaluation Results

A summary of the recall-precision results for the five main retrieval runs is shown for the French and English collections in Table 2. In each case, precision (P) values are given at five distinct recall (R) levels from 0.1 to 0.9. It may be seen from Table 2, that completely equivalent results are obtained for the thesaurus runs in English and French, thus confirming that equivalent language analysis tools can be built for both languages. In only three cases is the standard result not obtained (precision of about 0.5 and 0.35 for low and high recall values, respectively):

a) the French word form runs are not as good as the English ones;

b) the reduced thesaurus obtained by eliminating negative discrimination classes is not as good in English as in French.

In both cases, the explanation may be found in the greater variability of the French vocabulary. The French terminology is much less standardized than the English one in the area of documentation, and fewer word form matches are obtained as a result. The examples of Table 3 are cases in point: two matching words are found in English for "subject heading", and three for "information retrieval system"; these phrases are, however, translated into French in a variety of different ways, in each case providing fewer matching words. The French thesaurus, on the other hand, does reduce the variability of the word form terminology, producing results equivalent to those in English for the thesaurus runs.

When the negative discriminators included in Table 4 are removed from the thesaurus, the result improves for French but deteriorates in English, indicating that the sparser English vocabulary was originally translated into index vectors at the right level of exhaustivity to reflect the various aspects of document content. Whereas the more diversified French could stand a reduction by elimination of the poorer discriminators, a similar reduction evidently creates an English vocabulary that no longer adequately covers the document and query contact.

Two cross-language runs were made in which the French (English) documents, analyzed by the thesaurus method, were compared against the English (French) queries. The results of Table 5 indicate that the standard result is obtained for the English queries and French documents but not for the converse case. The reason is once again the variability of the French query vocabulary which does not get reduced to the English standard through the thesaurus transformation. Examples are shown in Tables 6 and 7 for queries 6, 7 and 13. The English word "retrieval" is seen to be rendered in French variously by "documentation", "recherche", and "récuperation". Each of the three French words occurs in a different thesaurus category (157, 280, and 330 respectively); in English a single category, number 330, applies in each case.

Thesauruses can, no doubt, be built which reduce the variability in any given language to that of some common standard. If this is done, different dictionaries and analysis tools appear to be required to operate within a single language, and between languages, respectively.

From an operational viewpoint, the old conclusion derived from earlier experiments is reinforced by the present results: document collections available for a given subject area in several natural languages can be processed fully automatically to produce substantially identical retrieval performances.

References

[1] G. Salton, Automatic Processing of Foreign Language Documents, Journal of the ASIS, Vol. 21, No. 3, May-June 1970, p. 187-194.

[2] B.R. Pevsner, A Comparative Evaluation of the Workings of the Russian and English versions of the "Empty-Nonempty 2" system, in The Automatic Translation of Texts, Moscow, October 1971.

[3] G. Salton, Experiments in Automatic Thesaurus Construction for Information Retrieval, Proc. IFIP Congress-71, Ljubljana, North Holland Publishing Co., Amsterdam, 1972

	English	French
Number of documents	52	52
Number of user queries	16	16
Number of distinct words in collection	1197	1340
Number of thesaurus classes	287	323
Number of entries (words) per thesaurus class	2.45	2.44

Thesaurus and Collection Statistics

Table 1

Query and Document Language	R	Word Form (standard) P	Word Form (D.V. weight) P	Thesaurus (standard) P	Thesaurus (D.V. weight) P	Thesaurus (Neg. Disc.) P
						*
	0.1	.5183	.4931	.5005	.4923	.4265
	0.3	.5183	.4931	.5005	.4923	.4265
English	0.5	.5183	.4931	.4796	.4715	.4109
	0.7	.3908	.4037	.3411	.3621	.3635
	0.9	.3908	.4037	.3411	.3621	.3635
		*	*			
	0.1	.4223	.4365	.5072	.4989	.5512
	0.3	.4223	.4365	.5072	.4989	.5512
French	0.5	.4015	.4365	.4655	.4989	.5512
	0.7	.3654	.4333	.3362	.3672	.4021
	0.9	.3654	.4333	.3362	.3672	.4021

*Standard result not obtained

Recall - Precision Results for English and French Language Material

Table 2

	French Documents English Queries	English Documents French Queries
Recall	Precision	Precision
		*
0.1	.5294	.4088
0.3	.5294	.4088
0.5	.5294	.4088
0.7	.3388	.3249
0.9	.3388	.3249

*Standard result not obtained

Comparison for Cross - Language Runs (Standard Thesaurus)

Table 5

Language	Query No.	Doc. No.	Query Entries	Document Entries	Matches
English	16	1	subject heading	subject heading	2
French	16	1	mots - vedettes	mots - clés	1
English	6	41	information retrieval system	information retrieval system	3
French	6	41	systèmes documentaires	systèmes de reperage d'information	1
English	8	13	library	library cards	1
French	8	13	bibliothèque	cartes de librairies	0

Differences in Word Matching Due to More Diversified French Vocabulary

Table 3

	French		English
✓	auteur	✓	author
✓	automate	✓	automatic
✓	bande		bibliographic
	calcul	✓	card
✓	carte	✓	catalog
✓	catalogage, catalogue	✓	computer
	classe	✓	development
	connexe, connexion	✓	different
	contenant, contenir, contenu	✓	document
	descripteur, descriptif, description	✓	give
✓	développe	✓	IBM
✓	difference, different	✓	index
✓	document	✓	information
✓	donne		library
✓	IBM	✓	machine
✓	index		need
✓	informatif, information		number
✓	machine		process
	mécanique	✓	program
	perforation, perforé		punch
	pratique	✓	research
✓	programme	✓	science, scientific
	rapport	✓	system
✓	recherche	✓	tape
✓	scientifique		type
✓	système, systematique		work
	technique		

English and French Negative Discriminators

Table 4

Query Number	English	French	Match
6	vocabulary of information retrieval	le vocabulaire de la documentation	No
	an understanding of information, storage, and retrieval	un tableau de la documentation, enregistrement et recherche	No
7	the amount of scientific publication in terms of analysis, control, storage and retrieval	la masse de publications scientifiques en termes d'analyse, d'emregistrement et de récuperation.	Yes

Thesaurus Category	English	French	
157	document	document	
	documentation	documentation	
280	project	recherche	
	research	projet	
	search		
333	extracted	extraction	
	extraction	récuperation	
	retrieval	récuperé	
		réperage	
		retrieval	
		retrouver	

Matching Samples for Cross-Language Thesaurus

Table 6

Query Number	Document Number	English Query	French Document	Match	French Query	English Document	Match
6	41	retrieval	réperage	yes	documentaire	retrieval	no
		information	information	yes	recherche	information	no
13	51	system	système	yes	recherche documentaire	system	no

Variability of French Query Vocabulary

Table 7

FRENCH

246-	1:	FASCICULE	246	
	2:	FEUILLES	246	
	3:	JOURNAUX	246	
	4:	NUMERO	246	
	5:	PERIODIQUE	246	
	6:	REVUE	246	178
247-	1:	JURIDICTION	247	
248-	1:	KWIC	248	
	2:	PERMUTATIONS	248	
	3:	PERMUTES	248	
250-	1:	COMPILATION	250	
251-	1:	LECTEUR	251	
	2:	LECTURE	251	
	3:	LIRE	251	
	4:	LISIBLE	251	
252-	1:	LEGALES	252	
253-	1:	LINEAIRE	253	
254-	1:	LIGNE	254	
	2:	RANGEES	254	
255-	1:	FLEXIONNELLE	255	
	2:	GRAMMATICAUX	255	
	3:	LANGAGE	255	400
	4:	LANGUE	255	400
	5:	LINGUISTIQUE	255	
	6:	MORPHOLOGIQUE	255	
	7:	SEMANTIQUE	255	
	8:	STRUCTURE	255	
	9:	SYNTAGMATIQUES	255	
	10:	SYNTAXE	255	
256-	1:	LISTE	256	
	2:	RECUEIL	256	
	3:	REPERTOIRE	256	
	4:	TABLES	256	

ENGLISH

246-	1:	ISSUE	246	
	2:	NEWSPAPERS	246	
	3:	PERIODICAL	246	
	4:	SHEET	246	
247-	1:	JURISDICTION	247	
248-	1:	KWIC	248	
250-	1:	COMPILATION	250	
	2:	COMPILED	250	
	3:	COMPILING	250	
251-	1:	READ	251	
252-	1:	LEGAL	252	
253-	1:	LINE	253	254
254-	1:	LINE	254	253
	2:	ROW	254	
255-	1:	GRAMMATICAL	255	
	2:	LANGUAGE	255	400
	3:	LINGUAL	255	400
	4:	LINGUISTIC	255	400
	5:	STRUCTURATED	255	
	6:	STRUCTURES	255	
	7:	STRUCTURING	255	
	8:	SYNTAX	255	
256-	1:	LIST	256	
	2:	TABLES	256	

English and French Thesaurus Excerpts

Figure 1