Why do Successful Search Systems Fail for Some Topics

Jacques Savoy
Computer Science Department, University of Neuchatel,
Rue Emile Argand 11, 2009 Neuchâtel, Switzerland
Jacques.Savoy@unine.ch

ABSTRACT
This paper describes and evaluates the vector-space and probabilistic IR models used to retrieve news articles from a corpus written in the French language. Based on three CLEF test-collections and 151 queries, we classify the poor retrieval results of difficult topics under 6 categories. The explanations we obtain from this analysis differ from those suggested *a priori* by our students. We use the Web to manually or automatically find related search terms to the original query. We evaluate these two query expansion strategies in order to improve mean average precision (MAP) and to reduce the number of topics for which no pertinent responses are listed among the top ten references returned.

Categories and Subject Descriptors
H.3.1 [**Content Analysis and Indexing**]: *Indexing methods; Linguistic processing*. H.3.3 [**Information Search and Retrieval**]: *Retrieval models*. H.3.4 [**Systems and Software**]: *Performance evaluation*.

General Terms
Measurement, Performance.

Keywords
Failure analysis, robust evaluation.

1. INTRODUCTION
During the last decade, the information retrieval (IR) domain has been confronted with larger volumes of information from which pertinent items must be retrieved in response to user requests. Given that Web users tend to submit short requests composed of only one or two words, new and effective IR models have been proposed to meet this difficult challenge. Some of these models include the Okapi model [1], language models [2], [3] or search strategies derived from the *Divergence from Randomness* [4] principle. Effective weighting formulae are suggested for these IR models in order to account for the following three aspects. First, it seems reasonable to assign more importance to terms appearing many times within a given document. Second, a wise search scheme must attribute, *ceteris paribus*, less weight to terms appearing in many documents. Third, the importance of documents containing many terms must be decreased.

Moreover, in order to increase the likelihood of obtaining a match between query terms and documents, search systems may also apply a stemming procedure to conflate word variants into a common stem. For example, when a query contains the word "cat," it seems reasonable to also retrieve documents containing the related word "cats," and this should also apply to those containing the terms "reliability" and "reliable." Finally, a stopword list may be used in order to eliminate words that appear relatively frequently in a document (such as "the", "your", "have", or "but"), given that they have no specific meaning.

Given these sound assumptions and procedures, it seems that *a priori* search engines should not fail to retrieve documents requested by users, especially when those retrieved share many of the same words contained in the request submitted.

When asked why an IR system might fail, computer science students taking an IR course (3rd year) submitted some of the following explanations. First, a search system will fail to discover pertinent documents if important information is missing. Second, if a request contains one (or more) spelling errors, the search engine will not be able to provide the expected answer. Third, if users introduce a single and ambiguous term such as "bank", it would not be really surprising if responses contain no relevant information. It is interesting to note that in commercial search engines such as Google, Yahoo or Wikipedia (free encyclopedia), the term "bank" only seems to refer to a financial institution. Fourth, the underlying weighting scheme may be deficient, placing too much importance on a given search term (or short part) of the query submitted. Finally, when asked if a query composed of four or more terms would be difficult for a search engine to process, the students argued that if a retrieved document contains at least 3 terms in common with the query, the information retrieved should most certainly be relevant to the submitted request. Note that in the rest of this paper, we will ignore this first explanation submitted, given that it usually results from an incorrect database.

During this experiment, it was clear to most participants that when submitting very short requests (one or two terms) polysemy words were a really important concern. Students felt that when users submitted three or more terms to a search engine, they would most likely (or even certainly) retrieve information that precisely matches their intentions and needs, and at least one pertinent item would appear on the list of the top ten retrieved documents. On the other hand, words characterized as synonyms (words appearing in the pertinent document but not having the same meaning such as "pizzeria," "restaurant," "snack bar," "coffee shop") are not clear and would be quickly recognized as potential problems.

In order to verify whether these given explanations correspond to reality, we used a relatively large collection of topics (151) extracted from the CLEF corpora. Based on twelve different search strategies, we identified *difficult* topics for all these IR models and discovered why these requests resulted in poor IR retrieval per-

formance. The rest of this paper is organized as follows. Section 2 describes the test collections used in our experiments while Section 3 presents an overview of all IR models evaluated in this study. An evaluation of twelve IR models and three query formulation is presented in Section 4, together with an analysis of the most difficult topics.

2. TEST-COLLECTIONS

To verify our assumptions, we used a French corpus created during the CLEF evaluation campaigns for the years 2001 [5], 2002 [6] and 2003 [7]. This document collection consists of articles extracted from the French newspaper *Le Monde* (1994), and from the Swiss news agency (*Schweizerische Depeschenagentur* or SDA, 1994-1995). Table 1 below lists some statistics.

Table 1. Selected statistics on our test-collections

	2001	2002	2003
Source	*Le Monde* 94 SDA 94	*Le Monde* 94 SDA 94	*Le Monde* 94 SDA 94 & 95
Size	243 MB	243 MB	331 MB
No. docs	87,191	87,191	129,806
Topics	#41 - #90	#91 - #140	#141 - #200

As the statistics show, the same documents appear in years 2001 and 2002. Moreover, these documents were written during the same year (1994) and cover news and events on political, economical, social, science and sport topics. The SDA corpus of course contains more documents on Swiss current affairs, while *Le Monde* includes more international coverage. During the year 2003, 42,615 documents extracted from the SDA corpus (for 1995) were added to the corpus. Topics 41 to 140 were searched in both the *Le Monde* and SDA 1994 corpuses, with the last 60 requests being taken the larger collection.

Based on the TREC format, each topic was structured into three logical sections comprising a brief title, a one-sentence description and a narrative specifying the relevance assessment criteria. In this study, we mostly used the shortest query formulation in order to reflect a more realistic search context. Queries were based on the topic's title section only and had a mean size of 2.91 search terms, while for their topic title and the descriptive parts, the mean query size was 7.51.

The available topics covered various subjects (e.g., "Pesticides in Baby Food," "Whale Reserve," "Renewable Power" or "French Nuclear Tests") and included both regional ("Swiss Initiative for the Alps") and international coverage ("Ship Collisions").

When inspecting the relevance assessments, we detected that nine topics (#64, #146, #160, #161, #166, #169, #172, #191 and #194) did not have any relevant items and thus the test set consisted of only 151 topics. When analyzing the number of relevant items per query, we found that the mean was 23.45 (median: 13, min: 1, max: 193 (for Topic #181) and standard deviation: 31.04).

3. IR MODELS

We used nine vector-space schemes and three probabilistic IR models in order to determine the really difficult topics. First we adopted a binary indexing scheme in which each document (or request) was represented by a set of keywords, without any weight. To measure similarities between documents and requests we computed the inner product (model denoted "doc=bnn, query=bnn" or "bnn-bnn"). We might take term occurrence frequency into account (or *tf*) with the corresponding retrieval model being denoted as "nnn-nnn". We might also take their inverse document frequencies (or *idf*) into account, applying cosine normalizations as described in [8].

Other variants might be created, especially given that the occurrence of a specific term in a document could be considered as a rare event. Thus, we might assign more importance to the first occurrence of a word, as compared to any successive, repeating occurrences. Therefore, the *tf* component would be computed as $\ln(tf) + 1.0$ (model denoted "ltc-ltc") or as $0.5 + 0.5 \cdot [tf / \max tf$ in a document]. Different weighting formulae might of course be used for documents and requests, leading to various weighting combinations. We might also consider that a term's presence in a shorter document would provide stronger evidence than it would in a longer document, leading to more complex IR models. Examples would be the IR models denoted by doc=Lnu [9] and doc=dtu [10] (see the Appendix for exact specifications).

In addition to these vector-space schemes, we also considered probabilistic models such as the Okapi [1] (BM25 weighting formula). As a second probabilistic approach, we implemented the DFR GL2 model, derived from the *Divergence from Randomness* principle [4] taht combined two information measures:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = (1 - Prob^1_{ij}) \cdot -\log_2[Prob^2_{ij}] \quad (1)$$

$$Prob^1_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad (2)$$
$$\text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((C \cdot mean\ dl) / l_i)]$$

$$Prob^2_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \quad (3)$$
$$\text{with } \lambda_j = tc_j / n$$

where w_{ij} indicates the indexing weight attached to term t_j in document D_i, l_i the number of indexing terms included in the representation of D_i, where tc_j represents the number of occurrences of term t_j in the collection, n the number of documents in the corpus, and C and *mean dl* are constants.

Finally we considered an approach based on a language model (LM) [2], [3], in which the underlying probability estimates are based on occurrence frequencies in document D and corpus C. Within this language model paradigm, various implementations [2] and smoothing methods [3] might be considered, and in this study we adopted a model proposed by Hiemstra [2] as described in Equation 4, which combines an estimate based on the document ($P[t_j | D_i]$, Equation 5) and on the corpus ($P[t_j | C]$, Equation 6).

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]] \quad (4)$$

$$\text{with } P[t_j | D_i] = tf_{ij}/l_i \quad (5)$$

$$\text{and } P[t_j | C] = df_j/lc \text{ with } lc = \sum_k df_k \quad (6)$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and *lc* the size of the corpus C.

Finally, during the indexing of requests and documents, all uppercase letters were replaced by their corresponding lowercase characters. All accents were removed, even though this process may accidentally conflate words having different meanings into the same form (e.g., the French word "tâche" (task) and "tache" (mark, spot)). The most frequent terms were also eliminated through applying a French stopword list (463 words). We applied a French suffix-stripping algorithm (see www.unine.ch/info/clef/)

that attempts to remove both inflectional (e.g., "cats" and "cat") and some derivational suffixes (e.g., "hopeful" and "hope").

4. EVALUATION

To measure retrieval performance, we adopted the mean average precision (MAP) computed by the TREC_EVAL program. To statistically determine whether or not one search strategy would be better than another, we applied the paired *t*-test. In this test, the null hypothesis H_0 states that both retrieval schemes produce similar performance. Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected (two-tailed test, significance level 5%).

4.1 IR Models Evaluation

Based on this methodology, Table 2 depicts the MAP resulting from three topic formulations: 1) title only (or T), 2) title and descriptive logical sections (TD), or 3) all three topic parts (TDN). The best performance under a given condition is shown in bold in this table. This experiment showed that the Okapi probabilistic model would result in the best retrieval performance, for all three topic formulations. The last lines of Table 2 list the mean, and also the mean obtained by considering the first seven IR models, which also performed best. When we increased query size from 2.91 terms (T) to 7.51 (TD), the retrieval effectiveness increased by 15.82%, and when using TDN, the enhancement was 25.91%, compared to the title-only queries.

It was also important to determine whether performance differences were really statistically significant. To do so we used the Okapi performance as a baseline, and our statistical test found that performance differences with the other IR models usually were statistical significant (denoted by an "*"). However, the difference between the Okapi and DFR GL2 models was not statistically significant when considering the TD (0.5074 vs. 0.4999) and TDN (0.5451 vs. 0.5410) topic formulations.

On the other hand, we wanted to verify whether increasing the query size would statistically improve the MAP. In this case, we considered the performance achieved under the T query formulation (second column of Table 2) as the baseline. Whenever our statistical test detected a statistically significant difference, we underlined the corresponding value. As depicted in Table 2, the TD and TDN query formulations statistically improved the MAP. The exceptions to this finding were the IR "nnn-nnn" and "bnn-bnn" models (which performed poorly compared to other search schemes).

It is known that the mean hides irregularities among observations, but it does have the advantage of summarizing a large number of values within a single number. In the current case and based on the shortest topic formulation (T), the Okapi model provided the best performance for only 37 queries out of 151 (or 24.5% of the observations). The resulting MAP was 0.4412, but the lowest performance level for this IR model was 0.0001 (achieved by Topic #200). The highest value was 1.0, obtained with 14 different requests (namely Topic #75, #83, #84, #105, #121, #123, #136, #144, #158, #173, #175, #188, #189 and #196). This high performance level was achieved by queries having one or two pertinent items, as shown in the first and second ranking positions in the output list. On the other hand for 21 requests, the Okapi system did not place one pertinent item in the top ten references. These observations illustrate that there may be a great deal of variability across topic evaluations. Users are not aware of mean performance given that they only see the resulting performance for their requests. They do however expect to obtain reasonable performance for all submitted requests (robust retrieval).

4.2 Failure Analysis

To understand why an effective search model such as the Okapi might fail to retrieve even one pertinent item, we analyzed the average precision (AP) for 151 queries. Additionally, in an effort to define really difficult topics, we considered all 11 IR models presented in Section 3 (evaluations listed in Table 2). Table 3 reports the AP of the most difficult or hard queries, defined as topics having null precision after 10 retrieved items (P@10). In these cases no pertinent information was retrieved and displayed to the user. This table lists the topic numbers, the best AP and rank of the first relevant item obtained by the best IR model, along with the model's name. Finally, reported in the last two

Table 2. Mean average precision for IR models with various topic formulations

	Mean average precision					
Query formulation	T	TD	TDN	T-manual	T-Google10	T-Yahoo10
Mean distinct terms / query	2.91	7.51	16.32	27.44	81.86	112.56
Model \ # of queries	151 queries	151 queries	151 queries	151 queries	151 queries	151 queries
doc=Okapi, query=npn	**0.4412**	**0.5074**	**0.5451**	**0.4414**	**0.4726**	**0.4718**
DFR GL2	0.4166*	0.4999	0.5410	0.4409	0.4715	0.4628
LM (λ=0.35)	0.3959*	0.4777*	0.5271*	0.4149*	0.3980*	0.3768*
doc=Lnu, query=ltc	0.3976*	0.4802*	0.5215*	0.4028*	0.3926*	0.3714*
doc=dtu, query=dtn	0.4066*	0.4714*	0.5104*	0.3837*	0.3751*	0.3481*
doc=atn, query=ntc	0.4062*	0.4590*	0.5063*	0.4181*	0.4363*	0.4313*
doc=ltn, query=ntc	0.3931*	0.4435*	0.4699*	0.4000*	0.3988*	0.3847*
doc=ntc, query=ntc	0.2791*	0.3248*	0.3575*	0.3042*	0.3183*	0.3008*
doc=ltc, query=ltc	0.2760*	0.3329*	0.3772*	0.3184*	0.3740*	0.3613*
doc=lnc, query=ltc	0.2842*	0.3580*	0.4160*	0.3331*	0.3879*	0.3885*
doc=nnn, query=nnn	0.1828*	0.1476*	0.1523*	0.1537*	0.1540*	0.1534*
doc=bnn, query=bnn	0.2074*	0.2024*	0.1389*	0.0491*	0.0287*	0.0238*
Mean	0.3314	0.3495	0.3977	0.3384	0.3507	0.3396
Mean over 7-best models	0.4027	0.4664	0.5070	0.4145	0.4207	0.4067
Improvement over T		+15.82%	+25.91%	+2.95%	+4.48%	+1.00%

columns is the difficulty rank of the corresponding topic when considering both the title and descriptive parts (TD) or all logical sections (TDN). A blank cell in the last two columns implies that P@10 was greater than 0 for the TD and TDN topics.

Table 3. The twelve most difficult topics (T formulation)

Topic	AP	Rank 1st rel. item	Best IR model	TD	TDN
#200	0.0002	817	Lnu-ltc		
#91	0.0017	56	dtu-dtn		
#155	0.0082	132	ntc-ntc	2	5
#156	0.0114	223	bnn-bnn	1	1
#120	0.0133	26	dtu-dtn	3	3
#48	0.0164	15	atn-ntc	5	6
#52	0.0181	34	atn-ntc	4	2
#151	0.0183	46	LM		
#117	0.0193	68	ntc-ntc		
#148	0.0344	41	atn-ntc	6	4
#51	0.0379	29	Lnu-ltc		
#109	0.0774	14	Okapi		

In an effort to explain the IR model's failure to list at least one pertinent item among the top ten, we might regroup the causes into two main groups: 1) system flaws (Category #1 to #3) and 2) topic intrinsic difficulties (Category #4 to #6).

Category 1: Stopword list. As a first explanation we found that the problem concerned letter normalization (uppercase replaced by a lowercase letter) and the use of a stopword list. Topic #91 ("AI en Amérique latine" and "AI in Latin America") is the second most difficult topic. With the T formulation, the best retrieval performance was 0.0017, achieved by the dtu-dtn scheme, with the first relevant article being ranked at 56. In this search system the query representation does not contain the acronym « AI » (Amnesty International) and thus without this information, the search system retrieved a lot of documents on « Latin America », without more specific thematic. In this case, the search system did not distinguish between the acronym "AI" (written in uppercase) and the verb form "ai" ("have" in French), a form included in the stopword list. In English we might encounter a similar problem with phrases such as "IT engineer," "vitamin a" or "US citizen". In the first case, the acronym could be mistaken for the pronoun "it" usually included in a stopword list, as is the case with the letter "a" and the pronoun "us". This example explains why commercial IR systems may not use a stopword list and thus index the documents under all available forms, and apply a stopword list only when analyzing the request [11]. As an additional problem, we should mention that relevant documents may cite a country (e.g., Mexico or Colombia) without explicitly linking the country name to the continent name.

Category 2: Stemming. The stemming procedure cannot always conflate all word variants into the same form or stem, as illustrated by Topic #117, ("Elections parlementaires européennes" and "European Parliament Elections"). Relevant articles have the terms "élections" and "européennes" or "Europe" with the query but using the noun form "parlement" instead of the adjective form "parlementaire" (parliamentary) as expressed in the topic. Although the search system was able to conflate the form "europe" and "européennes", it was not able to establish a link between the terms "parlement" and "parlementaire". As is shown with the English language in [12], a stemmer may remove too many endings and conflate words having quite different meanings under the same form.

Category 3: Spelling errors. The most difficult request was Topic #200 ("Innondationeurs en Hollande et en Allemagne" or "Flooding in Holland and Germany"). In an attempt to explain this poor performance, we should note that the term "Innondationeurs" is mispelled, and was the only request in our evaluation set that had this problem. This does not mean however that a spelling error should be viewed as a marginal problem. In the current case the correct spelling is "Innondations", as shown in the descriptive part. The query limited to "Holland and Germany" retrieved a lot of irrelevant material. In the descriptive part (TD) where the word "flooding" is correctly written, the request is no longer considered a "difficult" topic.

Category 4: Synonymy and language use. The third most difficult topic is Topic #155 ("Les risques du téléphone portable" and "Risks with Mobile Phones"), which illustrates how vocabulary can change across countries. For this request, the relevant documents used synonyms that are country dependant. In Switzerland, portable phones are usually called "natel", in Belgium "téléphone mobile", and "portable" In France. These country-dependant synonyms are not present in the descriptive and narrative parts of the topic and, as shown in Table 3, this request remains a difficult topic under TD or TDN formulations. The IR system included in its top ten results certain documents covering the use of portable phones in the mountains (and the risk of being in the mountains). Other retrieved articles simply presented certain aspects related to mobile phones (new joint ventures, new products, etc.).

Category 5: Missing specificity. A fifth failure explanation is found in Topic #156 ("Trade Unions in Europe"). The specific or desired meaning is not clearly specified or is too broad. This same difficulty occurs with Topic #120 ("Edouard Balladur"), Topic #48 ("Peace-Keeping Forces in Bosnia"), Topic #51 ("World Soccer Championship") or Topic #109 ("Computer Security"). With all these requests, the IR system listed top ranked articles having not one but at least two or three terms in common with the query. Placed at the top of the output list were short articles having all query terms in their title (or 3 out of 4 terms for Topic #48). For Topic #51 only, the additional information given in the descriptive part ("result of the final") really helped the IR system to list one pertinent article among the top ten results.

Category 6: Discrimination ability. The request retrieved a limited number of documents. For example in Topic #52, ("Dévaluation de la monnaie chinoise" and "Chinese Currency Devaluation") were used to retrieve information about the effects of devaluation. In this case, the three relevant articles had only one or two terms in common with the query. The terms "Chine" (also appearing in 1,090 other articles) and "monnaie" (currency, occurring in 2,475 documents) appeared in the first relevant document. In the second, only the noun "Chine" appeared to be in common with the topic's title, and in the last only the term "devaluation" (occurring also in 552 articles). The IR system therefore found it very difficult to discriminate between relevant and non-relevant documents, due to the fact that a lot of them had at least two terms in common with the query.

The same difficulty arose with Topic #151 ("Les sept merveilles du monde" and "Wonders of Ancient World"), and Topic #148 ("Dommages à la couche d'ozone" and "Damages in Ozone Layer").

Based on the shortest query formulation (T), we found six different reasons that might explain why the eleven search engines were able to display only one pertinent document in the top ten results. The only reason in common with the explanations in prior tests was "spelling error". Compared to previous work on English IR systems [13] that listed 9 possible failures, our list was shorter (6 reasons) and seems clearer. In [13] for example, it was not always clear under which category we should list the failure ("all systems emphasize one aspect, missing another *required* term" vs. "all systems emphasize one aspect; missing another aspect"). Moreover, the examples given in [13] were too short to make it possible to clearly understand all problems involved.

4.3 Expanding the Query

As shown in Table 3 (comparing performances between T and TD query formulations), the number of hard topics decreased from 12 to 6 when the user was able to provide more search terms. It would therefore be interesting to explore how we might improve retrieval effectiveness through expanding topic titles, especially for the queries shown in Table 3. This aspect was partly analyzed during the robust track at TREC in 2004 [14] or in 2005 [15]. Improving retrieval effectiveness seemed to be most promising among the approaches exploiting document collections (other than the evaluated corpus) or the Web to expand the original query. In this vein, Kwok *et al*. [16] suggested using word frequencies to automatically select the most appropriate terms extracted from the first 100 snippets to 40 full Web pages.

In our case, it might be wise to find related terms based on topic titles through submitting requests to newspaper Web sites (particularly those covering Swiss and French newspapers and agencies). These resources are not readily available, and also cultural, thematic and time differences may play a role in the effectiveness of such approaches [17]. We therefore decided to access the more valuable content available on the Web, more precisely that provided by the Google and Yahoo search engines.

In a first experiment, and contrary to previously proposed approaches that suggested fully automated query expansion, we expanded the queries manually. First, the topic title was given to the user who then submitted it as a request to Google. Second, based on the first page of results, the user was invited to expand the query by selecting the most appropriate references (note that only topic title were available to users). The modified queries were then submitted to all search systems, and the resulting MAP are shown in the fourth column of Table 2 (labeled "T-manual"). The mean query size increased from 2.91 distinct terms to 27.44. When observing users, we found that they usually added text snippets from the first three retrieved references. Compared to the T formulation, this manual query expansion method did not clearly improve MAP (we obtained similar MAP with the Okapi model, but it clearly showed significant improvement over the DFR GL2 model). When inspecting the hard topics, the expanded query formulation produced 9 queries with P@10 = 0 instead of 12. Topics #117 (stemming error), #151, #52 (both in the "discrimination ability") category and #51 were then able to retrieve one pertinent article in the first ten results.

In a second set of experiments, we sent the topic title to the Google and Yahoo search engines. The text snippets from the first ten references were automatically added to the original query, and this expanded request was then sent to all IR models. The performance achieved by these two automatic query expansion approaches are depicted in Table 2 under the label "T-Google10" and "T-Yahoo10". As shown in the second line, the number of distinct search terms increased from 2.91 to 81.86 and 112.56 search terms respectively. Comparing MAP for the Okapi and DFR GL2 models, we saw that retrieval performance increased significantly (e.g., for the DFR GL2 model, MAP increased from 0.4166 to 0.4715, or +13.1% using the Google-based expansion scheme). Moreover, the number of hard topics was reduced to 8 (Topics #151, #52, #117 and #155 disappeared from the "hard" set).

Finally as a blind query expansion technique we applied Rocchio's query expansion approach [9]. In this case we assumed that the top k ranked documents were relevant, without even examining them. From this set of documents we extracted the m most important terms added to the original query. In the current case based on the Okapi model, we found that the best value was k=3 articles and m=10 terms. With the T query formulation, the MAP achieved was 0.4412 and following Rocchio's blind-query expansion, the MAP increased to 0.4613, a statistically significant improvement. On the other hand the number of difficult topics increased from 12 to 18.

5. CONCLUSION

In this paper we evaluated nine vector-space IR schemes together with three probabilistic models in order to represent three of the most effective search paradigms: the Okapi [1], one of the *Divergence from Randomness* methods [4] and a language model [2] approach. Using a French test-collection created during three CLEF evaluation campaigns (151 queries), we found that the Okapi model performed best.

In a query-by-query analysis of the retrieval performance achieved by eleven IR models using the shortest topic formulation (title-only), we learned that the best approach did not always result in the best performance (only for 37 out of 151 requests, or 24.5%). Also, for 21 of these requests (14%) we were not able to find at least one pertinent item among the top ten references. As a result of this analysis we can categorize the IR failures under six main headings: 1) stopword lists, 2) stemming errors, 3) spelling errors, 4) synonymy and language use, 5) missing specificity, and 6) discrimination capability. It is worth noting that the top ranked but irrelevant documents found by all IR models had a large number of terms in common with the query, usually occurring nearby in a short context (e.g., in the document's title) and within a short article. With the exception of Category #3 (spelling error) no IR failure resulted from a single word.

To reduce the number of difficult topics (none of the IR systems were able to find a single relevant item in the first ten responses), we submitted requests to commercial search engines in order to find related terms, including both manual and automatic selections of text snippets. Query expansion approaches such as

these may in some circumstances improve MAP and reduce the number of hard topics. Moreover, these techniques seem to be more effective in reducing those difficult topics belonging to Category #2, #5 or #6.

ACKNOWLEDGMENTS

This research was supported by the Swiss NSF (under Grant #200020-103420).

6. REFERENCES

[1] Robertson, S.E., Walker, S., and Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *IP&M*, 36(1), 2000, 95-108.

[2] Hiemstra, D. *Using language models for information retrieval*. CTIT Ph.D. Thesis, 2000.

[3] Zhai, C., and Lafferty, J. A study of smoothing methods for language models applied to information retrieval. *ACM-TOIS*, 22(2), 2004, 179-214.

[4] Amati, G., and van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-TOIS*, 20(4), 2002, 357-389.

[5] Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds). *Evaluation of cross-language information retrieval*. LNCS #2406, Springer, Berlin, 2002.

[6] Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds). *Advances in cross-language information retrieval*. LNCS #2785, Springer, Berlin, 2003.

[7] Peters, C., Braschler, M., Gonzalo, J., and Kluck, M. (Eds). *Comparative evaluation of multilingual information access systems*. LNCS #3237, Springer, Berlin, 2004.

[8] Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. *IP&M*, 24(5), 19888, 513-523.

[9] Buckley, C., Singhal, A., Mitra, M., and Salton, G. New retrieval approaches using SMART. In *Proceedings of TREC-4*. Gaithersburg, MA, 1996, 25-48.

[10] Singhal, A., Choi, J., Hindle, D., Lewis, D.D. and Pereira, F. (1999). AT&T at TREC-7. In *Proceedings TREC-7*, Gaithersburg, MA, 1999, 239-251.

[11] Moulinier, I. Thomson Legal and Regulatory at NTCIR-4: Monolingual and pivot-language retrieval experiments. In *Proceedings NTCIR-4*, Tokyo, 2004, 158-165.

[12] Krovetz, R. Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR*. Pittsburgh, PA, 1993, 191-202.

[13] Buckely C. Why current IR engines fail. In *Proceedings ACM-SIGIR*, Sheffield, 2004, 584-585.

[14] Voorhees, E.M. Overview of the TREC 2004 robust retrieval track. In *Proceedings TREC-2004*. Gaithersburg, MA, 2004, 70-79.

[15] Voorhees, E.M. The TREC 2005 robust track. *ACM-SIGIR Forum*, 40(1), 2006, 41-48.

[16] Kwok, K.L., Grunfeld, L., Sun, H.L., and Deng, P. TREC2004 robust track experiments using PIRCS. In *Proceedings TREC-2004*. Gaithersburg, MA, 2004.

[17] Kwok, K.L., Grunfeld L., Dinstl, N., and Chan, M. TREC-9 cross-language, web and question-answering track experiments using PIRCS. In *Proceedings TREC-9*, Gaithersburg, MA, 2001, 417-426

7. APPENDIX

Table 4 shows the various indexing weights w_{ij} (for term t_j in a document D_i) formulas, where *n* indicates the number of documents, *t* the number of indexing terms, df_j the number of documents in which the term t_j appears, the document lengths (number of indexing terms) of D_i is denoted by nt_i, and *avdl*, *b*, k_1, *pivot* and *slope* are constants. For the Okapi weighting scheme, *K* represents the ratio of the length of D_i to l_i (sum of tf_{ij}), and the collection mean is noted by *avdl*.

Table 4. Weighting formulae

bnn	$w_{ij} = 1$	nnn	$w_{ij} = tf_{ij}$
ltn	$w_{ij} = (\ln(tf_{ij}) + 1) \cdot idf_j$	atn	$w_{ij} = idf_j \cdot [0.5 + 0.5 \cdot tf_{ij} / \max tf_{i.}]$
dtn	$w_{ij} = [\ln(\ln(tf_{ij}) + 1) + 1] \cdot idf_j$	npn	$w_{ij} = tf_{ij} \cdot \ln[(n-df_j) / df_j]$
Okapi	$w_{ij} = \left((k_1+1)\cdot tf_{ij}\right) \Big/ \left(K + tf_{ij}\right)$	Lnu	$w_{ij} = \dfrac{\left(1+\ln(tf_{ij}) \Big/ \ln(\text{mean tf})+1\right)}{(1-slope)\cdot pivot + slope \cdot nt_i}$
lnc	$w_{ij} = \dfrac{\ln(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t}(\ln(tf_{ik})+1)^2}}$	ntc	$w_{ij} = \dfrac{tf_{ij}\cdot idf_j}{\sqrt{\sum_{k=1}^{t}(tf_{ik}\cdot idf_k)^2}}$
ltc	$w_{ij} = \dfrac{(\ln(tf_{ij})+1)\cdot idf_j}{\sqrt{\sum_{k=1}^{t}((\ln(tf_{ik})+1)\cdot idf_k)^2}}$	dtu	$w_{ij} = \dfrac{(\ln(\ln(tf_{ij})+1)+1)\cdot idf_j}{(1-slope)\cdot pivot + slope \cdot nt_i}$