Scalable Multilingual Information Access

Paul McNamee and James Mayfield

Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road
Laurel MD 20723-6099 USA
{mcnamee,mayfield}@jhuapl.edu

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

3 September 2001

Outline

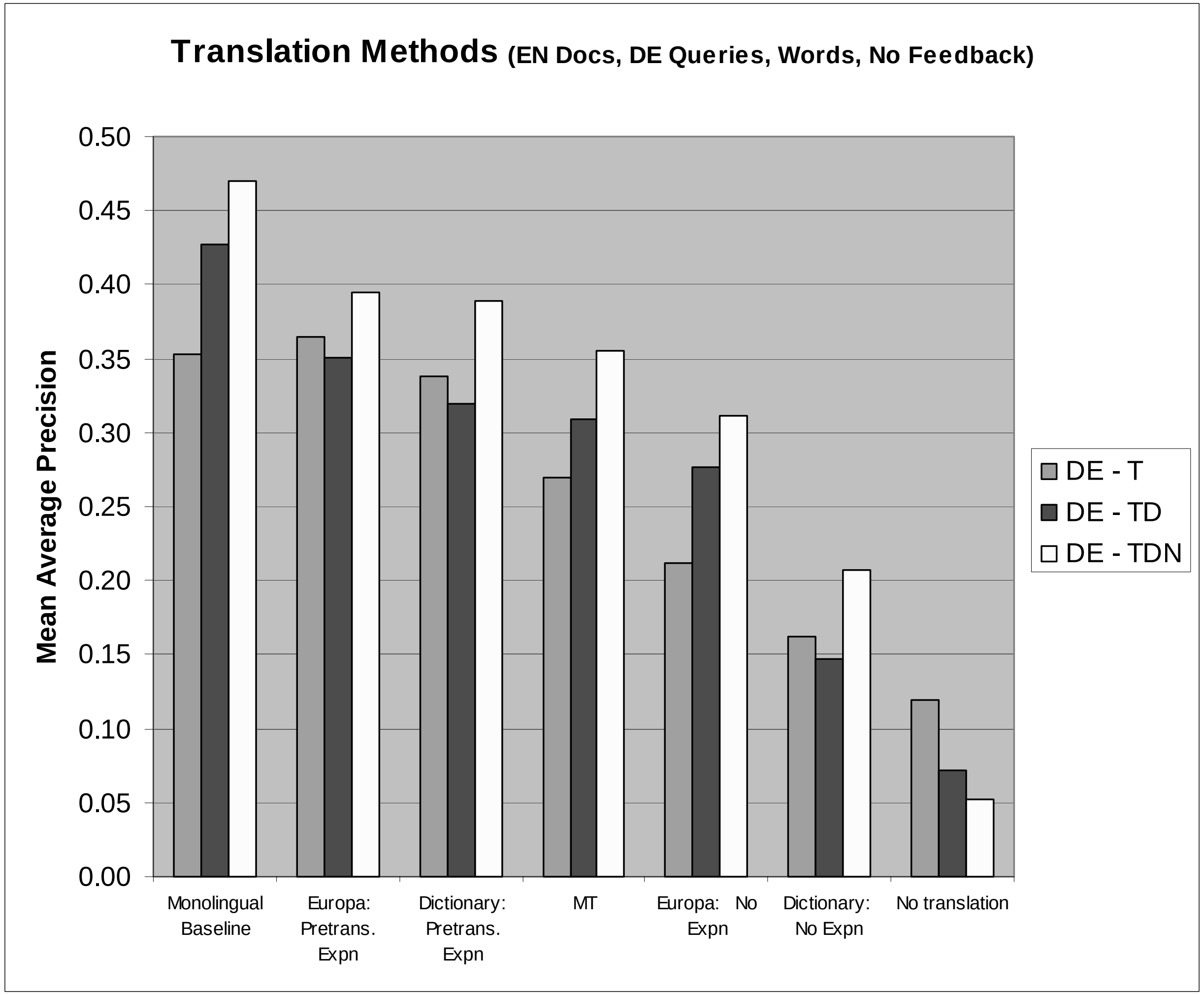
- Highlight work on 2001 collection
- Scalability
- CLEF-2002
  - Bilingual Retrieval
  - Multilingual Retrieval
- Conclusions and Future Work

RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Experiments on CLEF 2001 Test Set

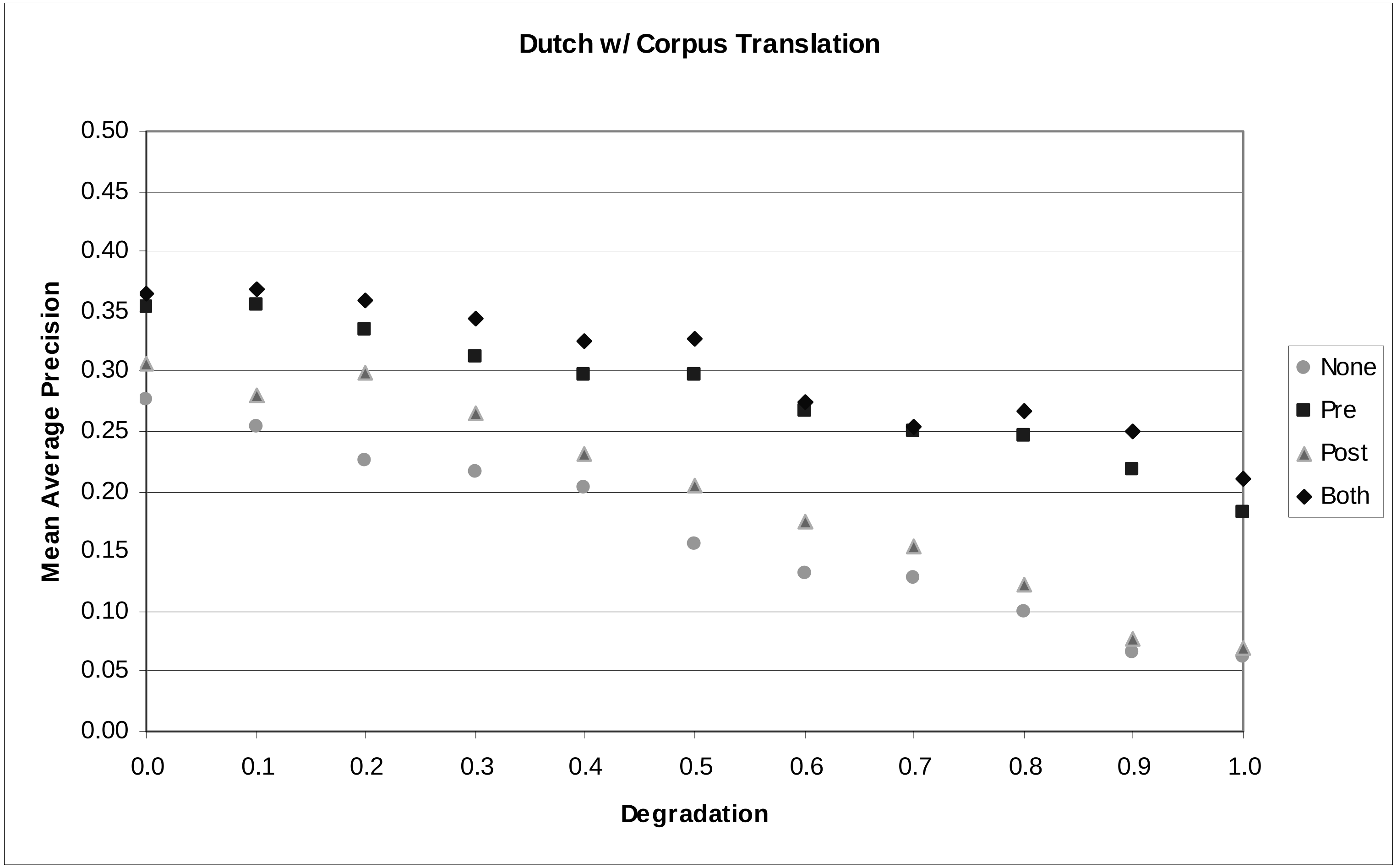
- **Comparison between different translation resources**
  - **Machine translation software, bidicts, aligned corpora, & simple cognate matching**
- **Investigation of query expansion techniques**
  - **Found that pre-translation expansion using comparable corpora is highly effective**
  - **Expansion mitigates losses due to poor resources**
- **Multilingual merging**
  - **Merge-by-rank and merge-by-score are comparable**

Rough Comparison of Translation Alternatives



JOHNS HOPKINS UNIVERSITY Applied Physics Laboratory

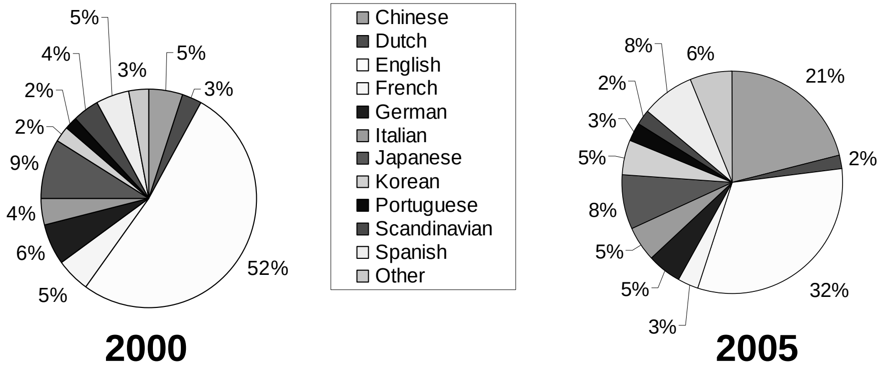
Effectiveness of Query Expansion Techniques



Scalability

- **Multilingual Information Access**
  - **Regardless of language**
- **Language-Neutral Methods are Attractive**
  - **Reduce human labor**
- **Conjecture: Software complexity over n-languages grows like O(n^k)**
  - **Therefore, we should reduce language-specific processing**

"68% of Internet users will be non-English speaking by 2005"
Global Reach, October 2000

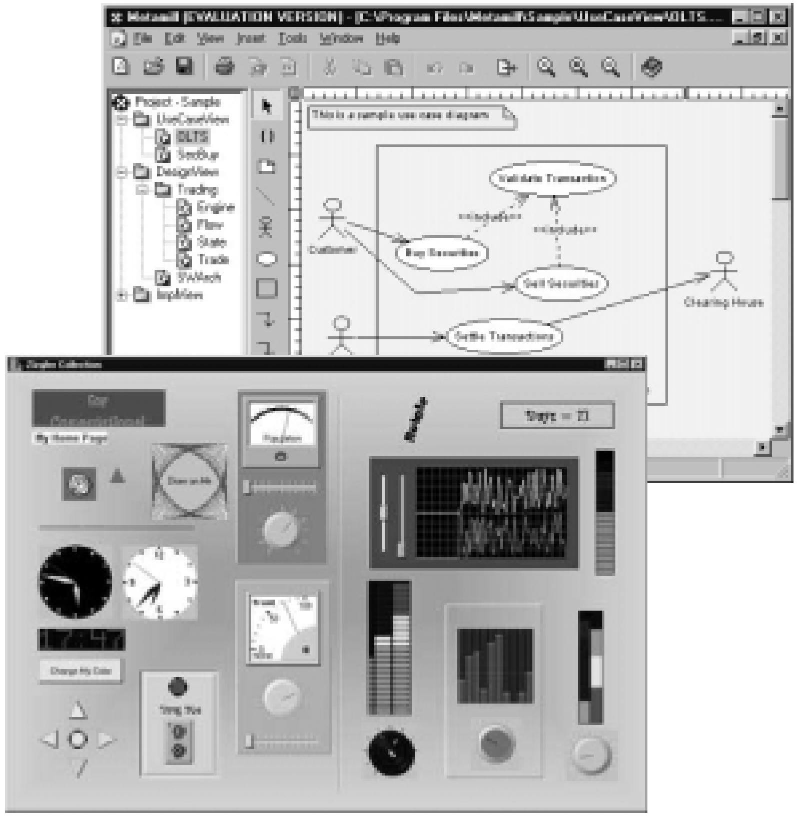


Computational Costs

- **The computer resources required for a CLIR application**
  - **Indexing the collection**
  - **Retrieval (and associated query-time processing)**
  - **Translation**
  - **Summarization & presentation of results**
- **Essentially CPU time, disk space, and memory**
  - **Compression is well-studied and commonly applied**
  - **Community has gravitated towards low-memory algorithms**
  - **Since disks and memory are cheap, time is the major concern**
- **Document translation for CLIR has been considered *too expensive***

Trend from SPECmarks to staff-months

- **Compiler products are now less concerned with optimal code generation**
  - ➢ **OOA&D support**
  - ➢ **Graphical components**
  - ➢ **Debugging**
  - ➢ **Profiling**
- **We might infer that developer time is more important than computer cycles (= user time)**
- **However, companies that buy compilers maximize profit by reducing developer costs, not user run-times**



Human Costs

- Two kinds of human costs required for a CLIR application

- End-users
  - Articulate a query (in one or more languages)
  - Sometimes assist in selecting query-translations
  - Might perform manual relevance feedback
  - Evaluate results
  - Extract information needed for current task

- System Developers
  - Assemble myriad non-standard resources
    - Stopword lists, stemmers, morphological analyzers, theasauri, phrase lists
    - Translation resources: dictionaries (in various formats), parallel corpora (which might need aligning), black-box MT software
  - Create index data structures
  - Write internationalized software

HAIRCUT Overview

- Hopkins Automated Information Retriever for Combing Unstructured Text
  - Statistical language model for retrieval
  - Supports large lexicons (useful for character n-grams)
  - Written in Java
    - Great high-level language
    - Native support for Unicode, multithreading
    - ‘Scalable’ if you own nice hardware
- Applied to CLIR tasks at TREC, CLEF, & NTCIR workshops
  - Language-neutral approach
  - Less is sometimes more

CLEF-2002 Approach

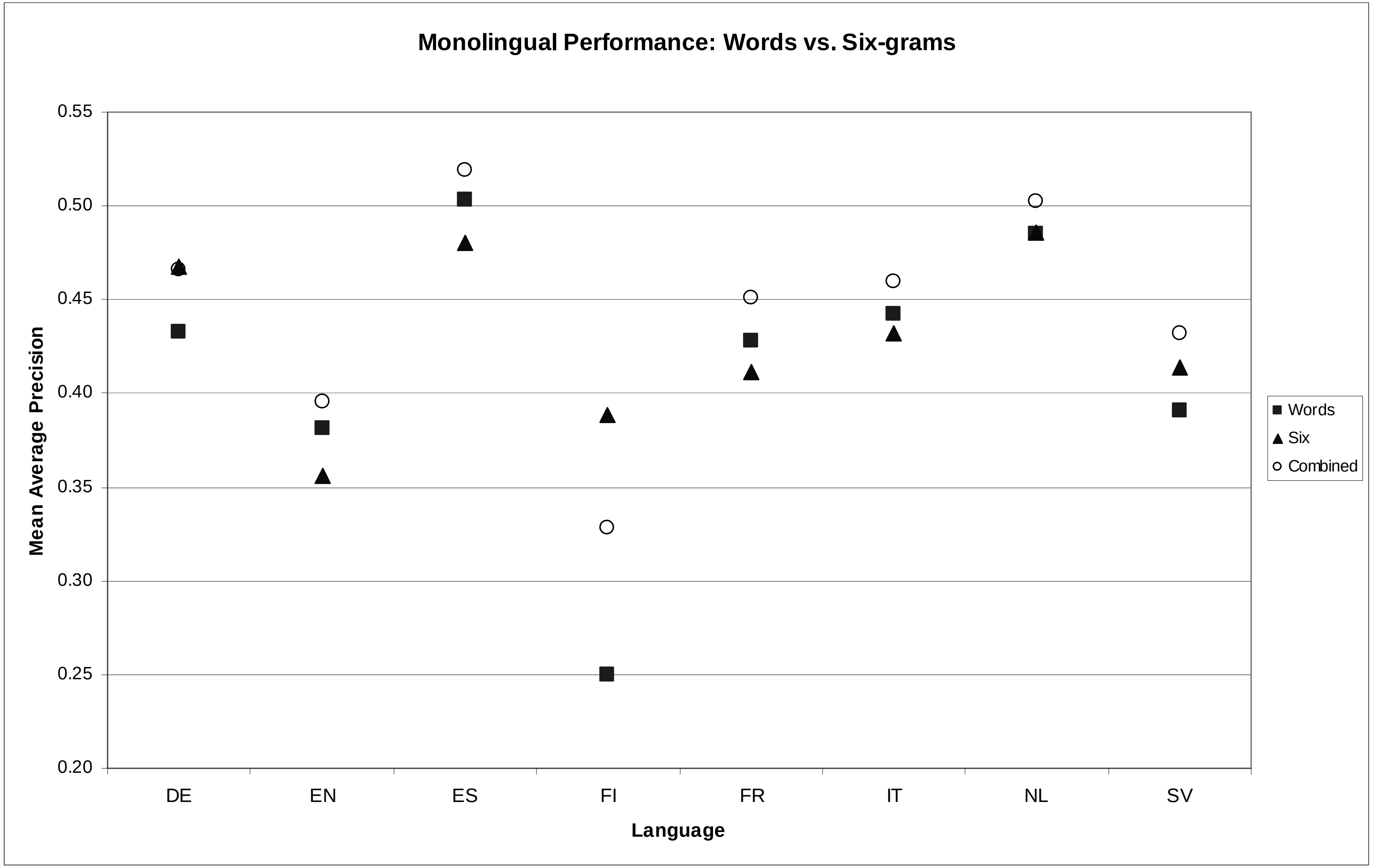
- **Monolingual Task**
  - **Two indexes per language: words & character 6-grams**
  - **Separate run-files were merged (by probability mass)**
- **Bilingual Task**
  - **Only used aligned corpus for translation and word-for-word translation; no use of n-grams**
  - **Pre-translation expansion performed using LA Times**
  - **Briefly looked at no-translation in close langauges**
- **Multilingual Task**
  - **Submitted runs using merge-by-rank and merge-by-score**
  - **Also examined translation of document representations**

For each task we only used the *title* and *desc* fields

Official Submissions

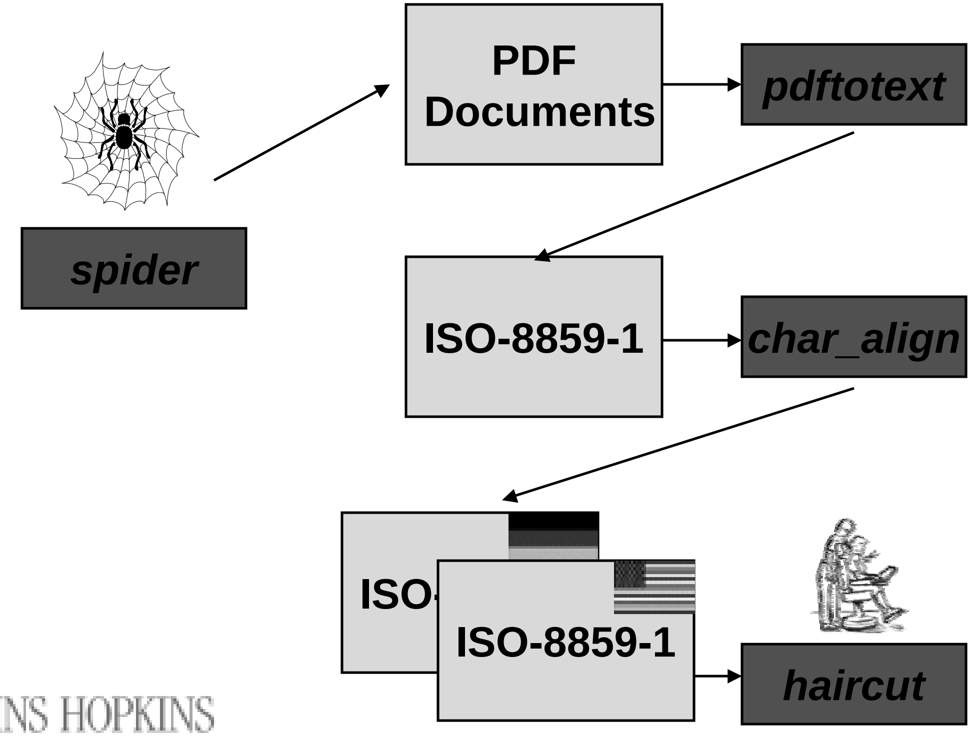
	Topic Fields	Average Precision	Precision at 5 docs	Recall at 1000	Relevant	# Topics
aplmode	TD	0.4663	0.5560	1792	1938	50
aplmoen*	TD	0.3957	0.5476	800	821	50
aplmoes	TD	0.5192	0.6120	2659	2854	50
aplmofi	TD	0.3280	0.3333	483	502	30
aplmofr	TD	0.4509	0.4800	1364	1383	50
aplmoit	TD	0.4599	0.5224	1039	1072	49
aplmonl	TD	0.5028	0.5960	1773	1862	50
aplmosv	TD	0.4317	0.4760	1155	1196	49

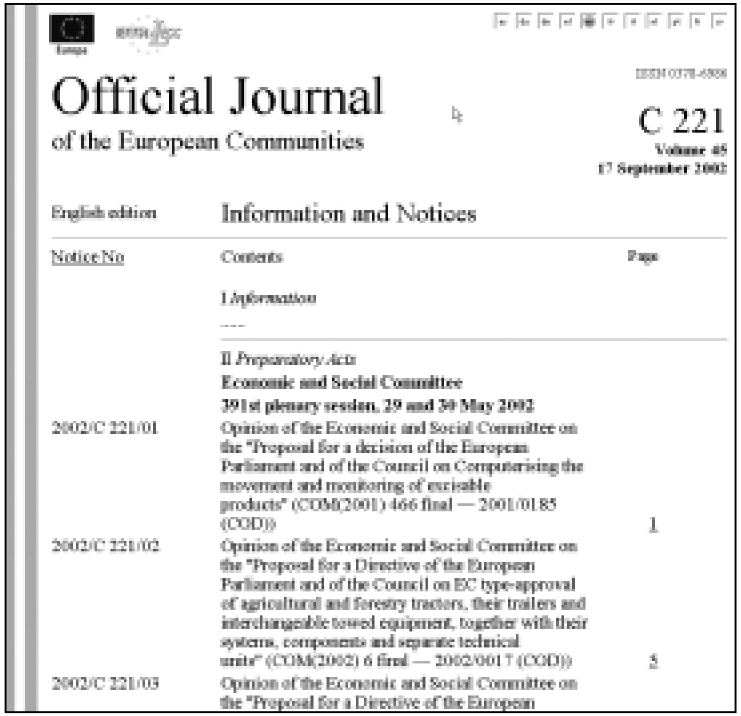
Comparing Indexing Terms by Language



Building a Translation Resource

- **Mined Official Journal of E.U.**
  - **Legal documents from http://europa.eu.int/**
  - **20GB of data obtained since 12/00 (200 MB / language)**
  - **Text in 11 languages produced as PDF**





Official Journal
of the European Communities

C 221
Volume 45
17 September 2002

English edition — Information and Notices

Notice No	Contents	Page
	I *Information*	

	II *Preparatory Acts*	
	Economic and Social Committee	
	391st plenary session, 29 and 30 May 2002	
2002/C 221/01	Opinion of the Economic and Social Committee on the "Proposal for a decision of the European Parliament and of the Council on Computerising the movement and monitoring of excisable products" (COM(2001) 466 final — 2001/0185 (COD))	1
2002/C 221/02	Opinion of the Economic and Social Committee on the "Proposal for a Directive of the European Parliament and of the Council on EC type-approval of agricultural and forestry tractors, their trailers and interchangeable towed equipment, together with their systems, components and separate technical units" (COM(2002) 6 final — 2002/0017 (COD))	5
2002/C 221/03	Opinion of the Economic and Social Committee on the "Proposal for a Directive of the European	

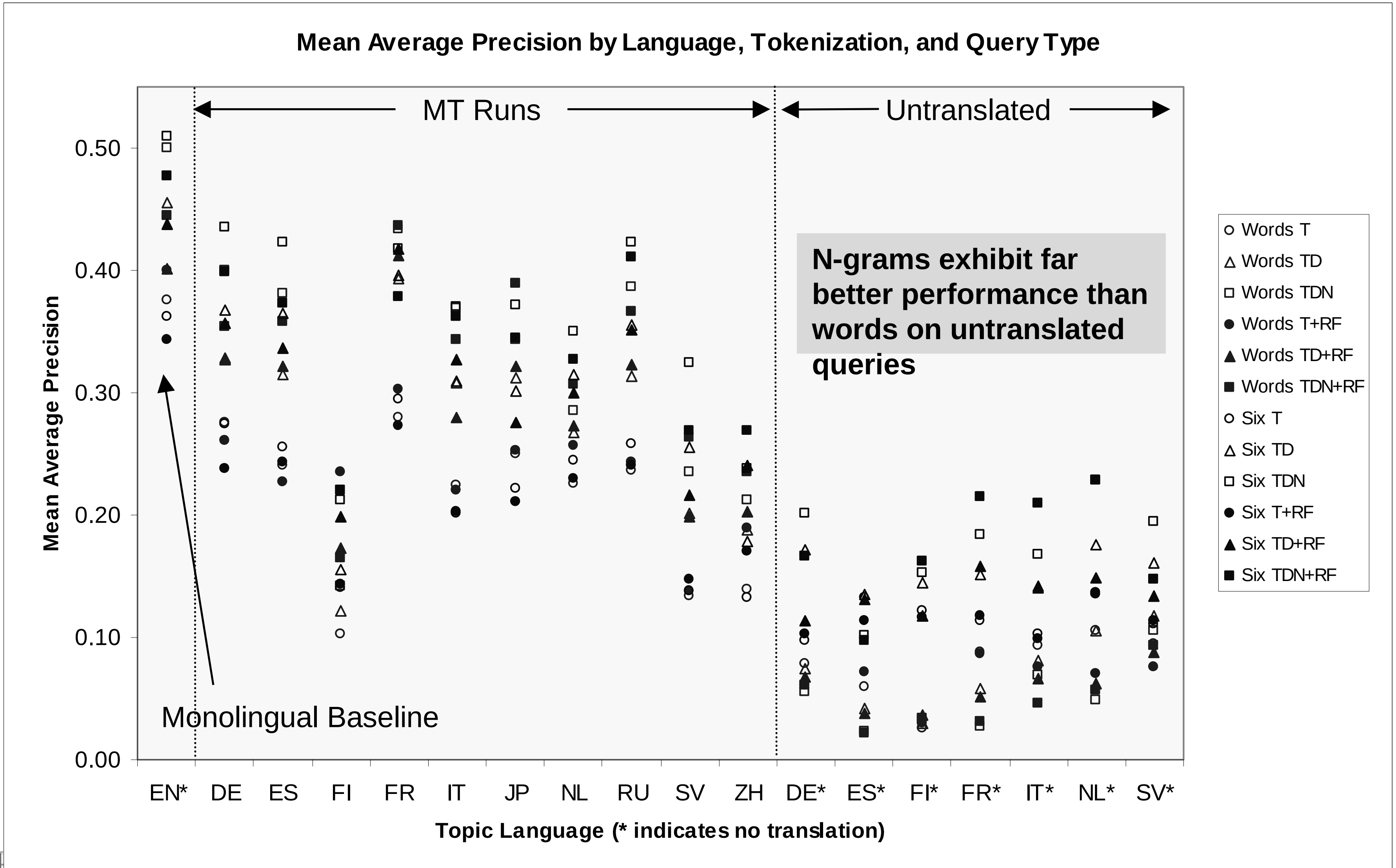
RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Bilingual Submissions

	Topic Fields	Average Precision	Precision at 5 docs	Recall at 1000	Relevant	# Topics
aplbiende	TD	0.3137	0.4160	1535	1938	50
aplbienes	TD	0.3602	0.4720	2326	2854	50
aplbienfi	TD	0.2003	0.2400	388	502	30
aplbienfr	TD	0.3505	0.4000	1275	1383	50
aplbienit	TD	0.2738	0.3347	934	1072	49
aplbiennl	TD	0.3516	0.3516	1625	1862	50
aplbiensv	TD	0.3003	0.4082	1052	1196	49
aplbipten	TD	0.4158	0.4857	753	821	42

English queries were expanded using the LA Times sub-collection. Then word-for-word query translation was performed using the single-best candidate translation extracted from the aligned corpus. With each language pair two runs were merged: one using pre-translation expansion alone, and one using both pre- and post-translation expansion.

CLEF-2001 MT vs. no translation



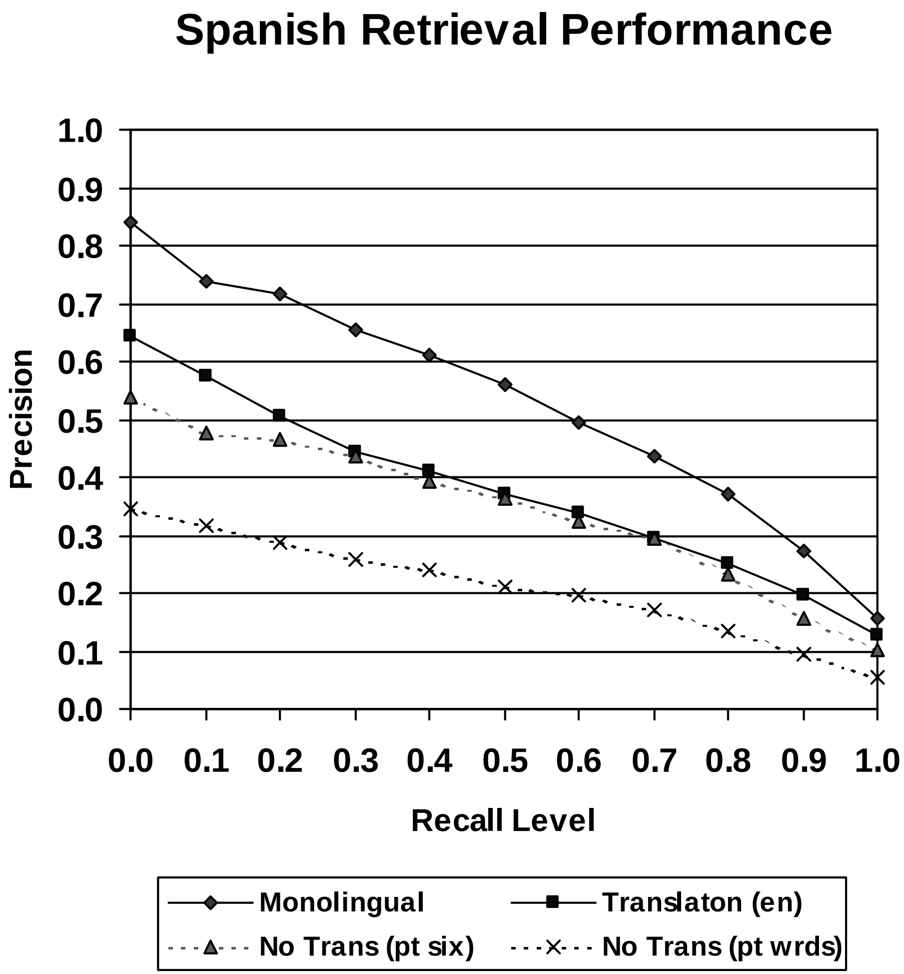
Without Any Translation

- **Direct translation may be infeasible between two given languages**
  - **Cognate matches can help in this scenario (Buckley et al. TREC-6; McNamee & Mayfield CLEF-2001; Shafer & Yarowsky – CoNLL-2002)**
- **We submitted a couple of runs using Portuguese topics to search Spanish documents**

	Fields	Term Type	Average Precision	Precison at 5 docs	Recall at 1000	# Rel
aplmoes	TD	words + n-grams	0.5192	0.6120	2659	2854
aplbienes	TD	words	0.3602	0.4720	2326	2854
aplbiptesa	TD	n-grams	0.3325	0.3920	2071	2854
aplbiptesb	TD	words	0.2000	0.2160	1589	2854

Portuguese-to-Spanish Results

- Can barely tell the difference between translated English queries and untranslated Portuguese queries
- Confirms that n-grams are more effective than unstemmed words for this scenario
- Previous work was restricted to retrieval of English documents



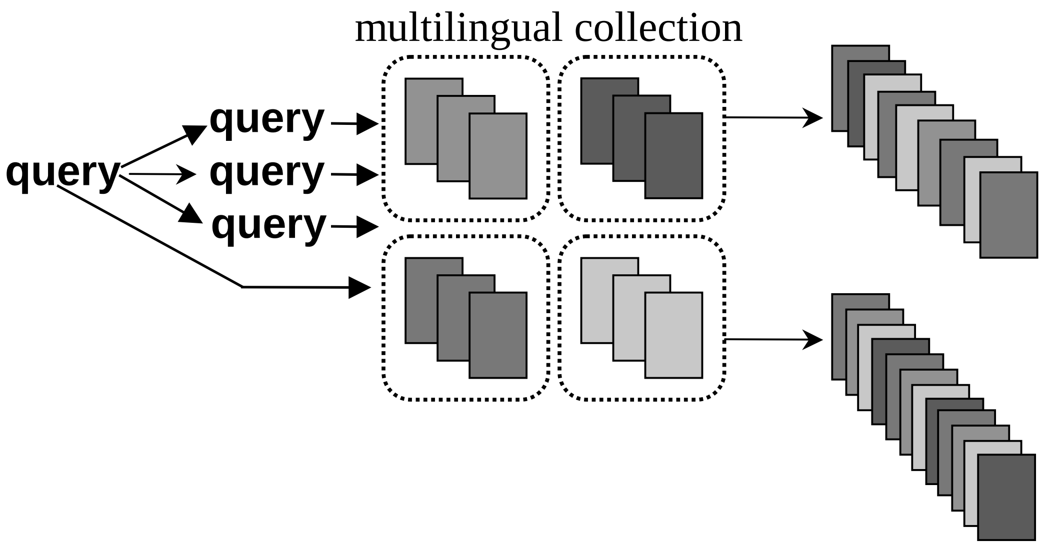
Multilingual Retrieval

- **In the multilingual problem, a single query language is used to search for relevant documents in multiple target languages**
  - **In many cases, relevant documents will be found predominently in a collection containing a particular language (non-uniform distribution)**
  - **It is more difficult to compare the relative relevance of documents in disparate languages than to rank documents in a single language**

- **Approaches**
  - **Distributed retrieval with merging**
  - **Unified collection (U. C. Berkeley in TREC-7, CLEF-2000)**
  - **Document Translation**

Distributed Retrieval & Merging

1. **Each language is separately indexed**
2. **Queries are translated from a single source language**
3. **The translated queries are run against the subcollections**
4. **The multiple ranked lists are combined**

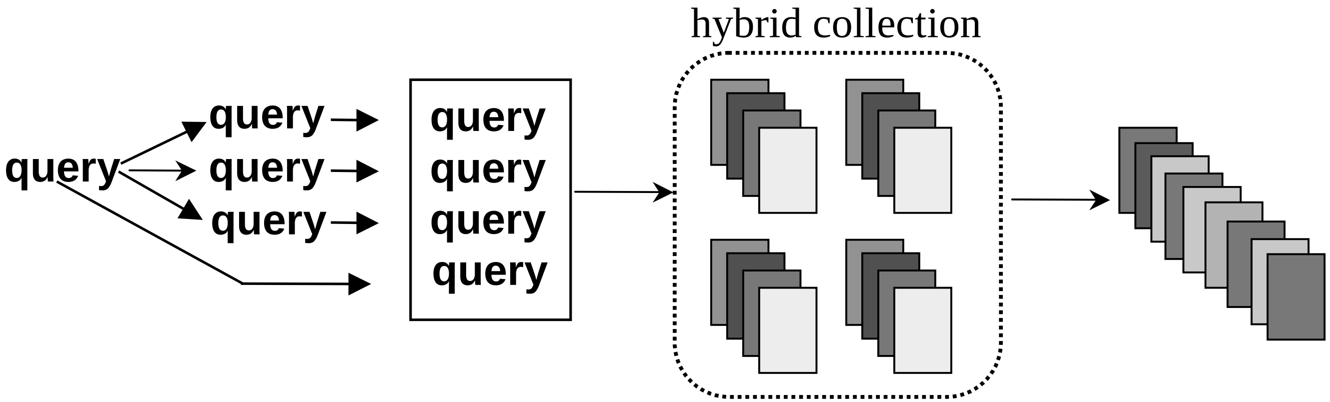


Merging by scores makes it possible to find the best documents regardless of language, but are scores really comparable?

Merge by rank (round-robin) is equitable, but may give undue consideration to languages with few relevant documents. Scalability is questionable when many, disparate languages are involved.

Unified Collection

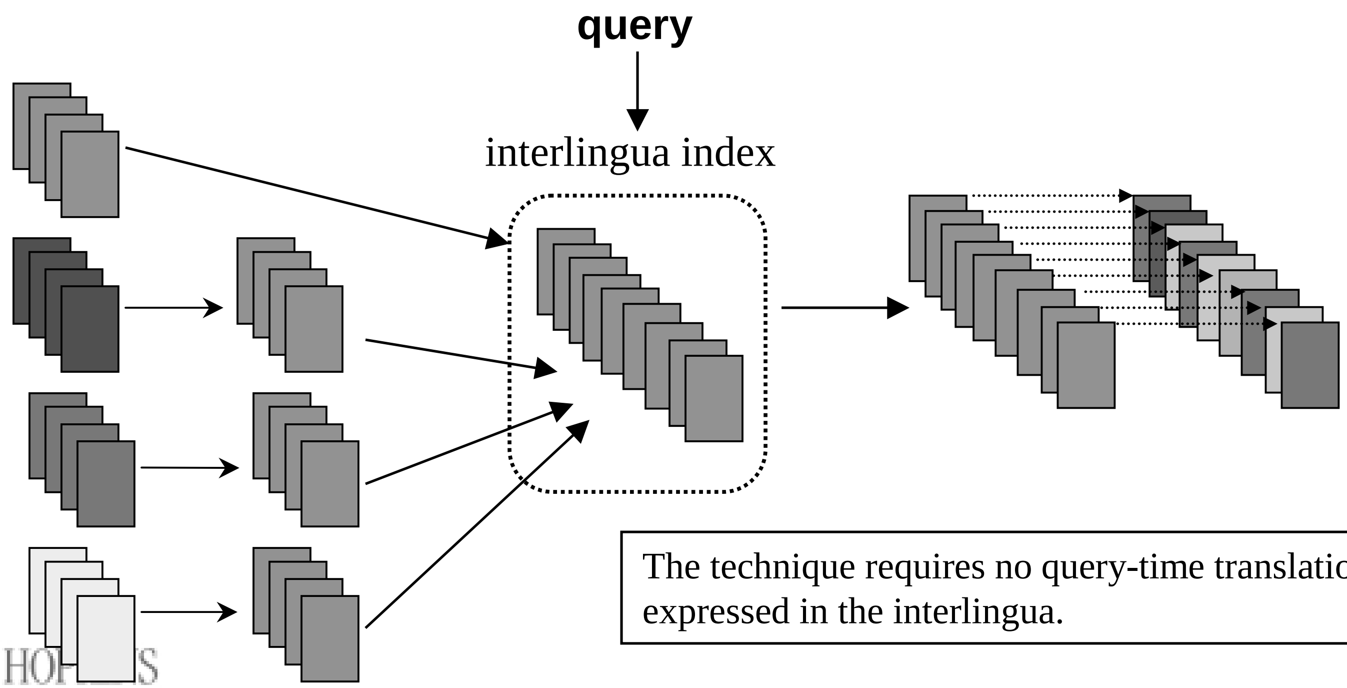
1. All documents are indexed in a common term-space
2. Queries are still translated from a single source language
3. A composite query is formed by combining translations
4. The single query is evaluated against the collection



Without word sense disambiguation, cognate matches should increase conflation; also, term statistics such as IDF will be somewhat altered compared to a monolingual collection. This technique does not require language identification

Document Translation (of sorts)

1. **All documents are indexed in their native language**
2. **The source language indexes are transduced into indexes using the term-space of an interlingua**
3. **The individual indexes are combined**
4. **Queries expressed in the interlingua are simply run against the new index**

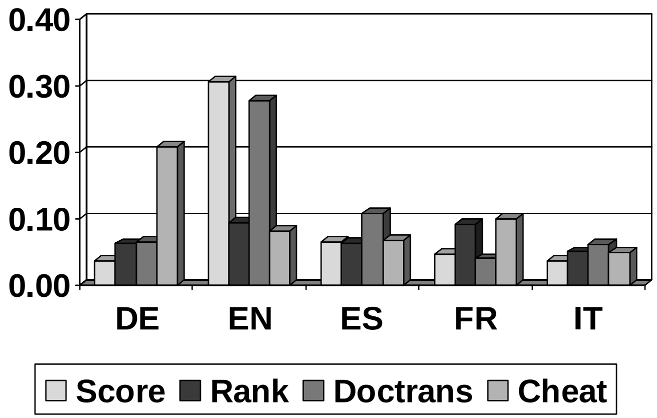


The technique requires no query-time translation, if queries are expressed in the interlingua.

Multilingual Submissions

	Query Lang.	Topic Fields	Average Precision	Prec. at 5 docs	Recall at 1000 (8068)	Comments
aplmuena	EN	TD	0.2070	0.4680	4729	Merge by score
aplmuenb	EN	TD	0.2082	0.4480	4660	Merge by rank
doctrans	EN	TD	0.2447	0.5760	3394	
doctrans + aplmuena	EN	TD	0.2456	0.5600	4766	Combine DT and QT
aplmucheat	ALL	TD	0.2265	0.4840	4772	Merged monolingual runs to isolate translation effects

MAP using Subcollection Qrels



RTDC
RESEARCH & TECHNOLOGY DEVELOPMENT CENTER

Issues with Document Transduction

- **Method for translation**
  - **Not FAHQMT. We did unbalanced word-to-words translation, preserving OOV words**
  - **Accomplished via an in-memory lookup table**
- **Less bias towards un-transduced sub-collection**
  - **'Translated' documents are larger and contain more noise**
- **Performance is good**
  - **Our implementation was less than 3x indexing time; can be reduced to a factor of 1.x**
  - **Provides a means of summarizing documents for speakers of the interlingua**
  - **18% improvement in mean average precision vs. merging**

Conclusions

- **Character n-grams and words comparable over many languages**
  - **6-grams clearly advantageous in Finnish**
- **Use of simple techniques (n-grams) can create problems**
  - **For example, using a dictionary for translation**
- **Document translation is viable and can be accomplished efficiently**
  - **Seems to outperform merge-by-rank and merge-by-score approaches to multilingual merging**

Future Work: Multilingual Filtering

- Nascent work to investigate text filtering over the CLEF test collections
- Operating under simple conditions
  - Split data temporally for training and testing
  - Assume pooled judgments from ad hoc evaluation are sufficient
  - Examining monolingual (many-language) filtering and cross-language filtering
- Interested in talking with others interested in this problem

Statistical Language Model for Retrieval

- **HAIRCUT uses a linguistically-motivated probabilistic model to estimate the probability that a document is relevant given a query**
  - **Hiemstra and de Vries, (*CTIT Tech. Report*, May 2000)**
  - **Miller, Leek, and Schwartz, *(SIGIR-99,* August 1999)**

Q = query
q = word in query
D = document
R = set of relevant documents
λ = a random Boolean variable

$$P(D \in R \mid Q) = \frac{P(Q \mid D \in R) P(D \in R)}{P(Q)} \quad \textit{Bayes law}$$

$$\propto P(Q \mid D \in R) \quad \textit{assume constant priors}$$

$$= \prod_{q \in Q} P(q \mid D \in R) \quad \textit{Naïve Bayes assumption}$$

$$= \prod_{q \in Q} \left[P(q \mid D \in R, \lambda) P(\lambda) + P(q \mid D \in R, \bar{\lambda}) P(\bar{\lambda}) \right] \quad \textit{introduce } \lambda$$

$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R, \lambda) + (1-\alpha) P(q \mid D \in R, \bar{\lambda}) \right] \quad \textit{define } \alpha = P(\lambda)$$

$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R, \lambda) + (1-\alpha) P(q \mid \bar{\lambda}) \right] \quad \textit{if q ind. of D given } \lambda$$

$$= \prod_{q \in Q} \left[\alpha P(q \mid D \in R) + (1-\alpha) P(q) \right] \quad \textit{because lambdas are ugly}$$

relative document term frequency

mean relative document term frequency

Default values for alpha:

0.30 words
0.15 6-grams

Using a fixed value for alpha works empirically, but can we do better?

IDF-like effect occurs due to the contribution from the 'generic language' probability (mean relative document term frequency).