Exeter at CLEF 2002: Using English for non-English Bilingual Retrieval

Adenike M. Lam-Adesina **Gareth J. F. Jones**

Department of Computer Science

University of Exeter, Exeter, U.K.

Main Objective of Participation

Explore effectiveness of monolingual and bilingual retrieval for non-English documents using only English retrieval.

Overview

- Non-English documents and topics translated into English using commercial machine translation.
  - Documents translated using *SYSTRAN 3.0.*
  - Topics translated separately using *SYSTRAN 3.0* and *Globalink Power Translator Pro 6.4*.

CLIR Strategy

- Retrieval using the City University research distribution of the Okapi system.
  - Around 260 stop words removed from the texts, Porter stemming applied and a small set of synonyms used.
- Okapi augmented with summary-based pseudo relevance feedback (Lam-Adesina & Jones SIGIR 2001).
- Pilot searching incorporating TREC-8 ad hoc document collection.
- Investigate merging topic translations.

Experimental Results

Experiments for Italian and Spanish document collections.

All runs use Topic Title and Description fields.

	Topic Language				
Av. Prec.	English	Italian	French	German	Spanish
Baseline	0.330	0.388	0.324	0.293	0.319
TC $ow(i)$, TC $cfw(i)$	0.374	0.453	0.375	0.341	0.363
PC $ow(i)$, TC $cfw(i)$	0.407	0.414	0.373	0.347	0.362
PC $ow(i)$, PC $cfw(i)$	0.399	0.376	0.365	0.335	0.333
CC $ow(i)$, CC $cfw(i)$	0.406	0.442	0.371	0.346	0.371

Retrieval results for Italian documents with Systran topic translation.

Experimental Results

	Topic Language				
Av. Prec.	English	Italian	French	German	Spanish
Baseline	0. 330	0.388	0.310	0.305	0.337
TC $ow(i)$, TC $cfw(i)$	0.374	0.394	0.358	0.377	0.371
PC $ow(i)$, TC $cfw(i)$	0.407	0.378	0.359	0.364	0.379
PC $ow(i)$, PC $cfw(i)$	0.399	0.375	0.352	0.325	0.360
CC $ow(i)$, CC $cfw(i)$	0.406	0.384	0.366	0.372	0.378

Retrieval results for Italian documents with PTP topic translation.

Experimental Results

	Topic Language				
Av. Prec.	English	Spanish	French	German	Italian
Baseline	0. 371	0.442	0.357	0.298	0.331
TC $ow(i)$, TC $cfw(i)$	0.412	0.475	0.382	0.318	0.359
PC $ow(i)$, TC $cfw(i)$	0.426	0.473	0.390	0.329	0.369
PC $ow(i)$, PC $cfw(i)$	0.420	0.420	0.373	0.334	0.342
CC $ow(i)$, CC $cfw(i)$	0.443	0.490	0.383	0.358	0.345

Retrieval results for Spanish documents with Systran topic translation.

Experimental Results

	Topic Language				
Av. Prec.	English	Spanish	French	German	Italian
Baseline	0.371	0.419	0.377	0.340	0.339
TC $ow(i)$, TC $cfw(i)$	0.412	0.445	0.414	0.387	0.369
PC $ow(i)$, TC $cfw(i)$	0.426	0.466	0.411	0.379	0.399
PC $ow(i)$, PC $cfw(i)$	0.420	0.431	0.396	0.350	0.370
CC $ow(i)$, CC $cfw(i)$	0.443	0.478	0.407	0.392	0.407

Retrieval results for Spanish documents with PTP topic translation.

Observations

- Monolingual retrieval is best in both cases despite translation of both documents and topics.
- Pilot searching using only the English document pilot collection is generally poor.
- Effectiveness of pilot searching appears to be related to quality of document translations.
  - Systran Spanish document translation generally better than Italian document translation.
- Merged pilot collection is better, but its effectiveness is still related to quality of document translations.

Experimental Results: Query Merging

	Topic Language				
Av. Prec.	English	Italian	French	German	Spanish
PC $ow(i)$, TC $cfw(i)$, Sys	0.330	0.414	0.373	0.347	0.362
PC $ow(i)$, TC $cfw(i)$, PTP	0.330	0.378	0.359	0.364	0.379
PC $ow(i)$, TC $cfw(i)$, QM	—	0.421	0.348	0.377	0.373
PC $ow(i)$, TC $cfw(i)$, QM2	—	0.411	0.377	0.368	0.365

Query merging retrieval results for Italian documents.

Experimental Results: Query Merging

	Topic Language				
Av. Prec.	English	Spanish	French	German	Italian
PC $ow(i)$, TC $cfw(i)$, Sys	0.426	0.473	0.390	0.329	0.369
PC $ow(i)$, TC $cfw(i)$, PTP	0.426	0.366	0.411	0.379	0.399
PC $ow(i)$, TC $cfw(i)$, QM	—	0.470	0.419	0.359	0.354
PC $ow(i)$, TC $cfw(i)$, QM2	—	0.468	414	0.354	0.379

Query merging retrieval results for Spanish documents.

Further Work

- Investigate data fusion merging.
- Translate French and German collections and repeat study.
  - Also run with English collection to compare behaviour.
- Multilingual retrieval experiments.