Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection

Gareth J. F. Jones, Declan Groves, Anna Khasin, Adenike Lam-Adesina,
Bart Mellebeek. Andy Way
School of Computing, Dublin City University, Dublin 9, Ireland
email: {*gjones,dgroves,akhasin,adenike,bart.mellebeek,away*}*@computing.dcu.ie*

Abstract

For the CLEF 2004 ImageCLEF St Andrew's Collection task the Dublin City University group carried out three sets of experiments. We carried out standard cross-language information retrieval (CLIR) runs using topic translation using machine translation (MT), combination of this run with image matching results from the VIPER system, and a novel document rescoring approach based on automatic MT evaluation metrics. Our standard CLIR approaches works well in comparison on this task. Encouragingly combination with image matching lists can produce small positive changes in the overall retrieval output. However, rescoring using the MT evaluation metrics in their current form significantly reduces retrieval effectiveness.

1 Introduction

Dublin City University's participation in the CLEF 2004 ImageClef St Andrew's task comprised three sets of experiments for Dutch, French, German, Italian and Spanish topic languages. First, we explored the application of our existing CLIR system used in previous CLEF workshops [1] with topic translation using three web-based translation resources. Second, combined this with the image matching results resulted provided by the track organisers' generated using the VIPER system. Finally, we explored a novel approach to rescoring the potentially relevant documents retrieved using our standard system based on automatic machine translation (MT) evaluation metrics.

The paper briefly outlines the details of our retrieval system before giving results for the first two sets of experiments. A separate section describes the MT evaluation metrics and gives results.

2 Standard CLIR Methodology

2.1 Retrieval System

The basis of our experimental retrieval system is the City University research distribution version of the Okapi system, as used in our previous CLEF participation [1]. The documents and search topics were processed to remove stopwords from a list of about 260 words; suffix stripped using the Okapi implementation of Porter stemming and terms were indexed using a small set of synonyms.

Terms are weighted using the standard BM25 weighting scheme and all runs use our summary-based pseudo relevance feedback (PRF) method [2]. The summary generation method combines Luhn's keyword cluster method, a title terms frequency method, a location/header method and a query-bias method from to form an overall significance score for each sentence.

	Dutch	French	German	Italian	Spanish
Prec. 10 docs	0.424	0.488	0.536	0.396	0.416
Av Precision	0.384	0.427	0.464	0.402	0.383
Rel. Ret.	698	631	695	606	654

Table 1: Baseline retrieval runs using Systran topic translation.

		SDL	INT	ST	MG
Dutch	Prec. 10 docs	0.472	0.276	0.500	0.472
	Av Precision	0.398	0.273	0.432	0.421
	Rel. Ret.	683	637	709	791
French	Prec. 10 docs	0.472	0.532	0.496	0.432
	Av Precision	0.409	0.466	0.431	0.399
	Rel. Ret.	666	707	658	695
German	Prec. 10 docs	0.592	0.528	0.540	0.632
	Av Precision	0.501	0.468	0.474	0.531
	Rel. Ret.	763	804	691	804
Italian	Prec. 10 docs	0.400	0.288	0.444	0.384
	Av Precision	0.366	0.288	0.438	0.351
	Rel. Ret.	633	591	602	639
Spanish	Prec. 10 docs	0.484	0.316	0.460	0.448
	Av Precision	0.444	0.318	0.406	0.398
	Rel. Ret.	767	666	649	755
Spanish (rev.)	Prec. 10 docs	0.532	0.320	0.492	0.488
	Av Precision	0.472	0.312	0.410	0.446
	Rel. Ret.	775	657	647	774

Table 2: Retrieval runs with PRF.

2.2 Experimental Results

For the all the experiments report here the Okapi parameters were set using the provided training topics as follows; $K1 = 1.0$ and $b = 0.5$ for baseline runs and $K1 = 1.5$ and $b = 0.6$ for PRF runs. The 20 top ranked PRF expansion terms were selected from the summaries of the top 5 ranked documents. The original topic terms were upweighted by a factor of 3.5 relative to terms introduced by PRF. There are a total of 829 relevant images available in the collection.

Topics were translated into English, the document language, using the following web-based MT systems: Systran (`http://www.systransoft.com/`), SDL (`http://www.freetranslation.com/`) and InterTrans (`http://www.intertrans.com/`). Results are show for CLIR using each separate translation and a term union merged translation.

Baseline Runs Table 1 shows baseline retrieval runs for Systran without application of PRF. Results in all languages appear reasonable with little apparent correlation between precision and recall figures.

Feedback Runs The text annotations of the images are typically very short, typically comprising only a few sentences. In developing our system for the PRF compared our summary-based approach developed for use with newspaper archives with a standard PRF approach selecting terms from complete documents. We were a little surprised to find that selecting terms from summaries of even these short documents worked better on the development topics than the whole document approach.

Table 2 shows feedback results for each topic language with the three MT systems and the merged translated topics. Separate results are shown for the original and later released revised Spanish topics. Comparing all these runs we can see that for Systran, PRF on on average produces

		SDL	INT	ST	MG
Dutch	Prec. 10 docs	0.480	0.276	0.508	0.464
	Av Precision	0.394	0.273	0.433	0.419
	Rel. Ret.	638	637	709	791
French	Prec. 10 docs	0.472	0.520	0.496	0.428
	Av Precision	0.407	0.466	0.428	0.399
	Rel. Ret.	666	707	658	695
German	Prec. 10 docs	0.604	0.524	0.548	0.636
	Av Precision	0.501	0.467	0.474	0.532
	Rel. Ret.	763	804	691	804
Italian	Prec. 10 docs	0.400	0.288	0.440	0.392
	Av Precision	0.369	0.289	0.437	0.351
	Rel. Ret.	633	591	602	639
Spanish	Prec. 10 docs	0.472	0.324	0.452	0.444
	Av Precision	0.441	0.316	0.405	0.397
	Rel. Ret.	767	666	649	755

Table 3: Retrieval runs fusing PRF runs with standard VIPER image matching results.

an improvement in average precision for each language pair, although there is no clear trend for relevant document recall. Comparing between the alternative topic translations it can be seen that different systems on average produce the best average precision for different language pairs, although in general InterTrans is the least effective. Results for the merged topics are rather mixed. It was hoped that the increased term coverage would improve recall and aid precision; this does happen in some cases, but in others it reduces effectiveness. Further investigation is needed into specific success and failures to see if any general conclusions can be made.

Text and Image Combination Runs Table 3 shows results for a simple sum combination of the matching score for the PRF runs shown in Table 2 and the provided VIPER runs. These results are only slightly different from the PRF runs. However, some potentially important positives can be taken from this. First, in image retrieval it is often found that adding image matching information cannot improve over text caption only retrieval. For our experiments in some cases the image matching score does help, albeit only marginally. Second, the VIPER system was not adjusted for the St Andrew's collection task, suggesting that a better image matching run should be possible with some task specific training of the image matching process.

3 Machine Translation Quality Metric Runs

In recent years, several automatic MT evaluation methods have been proposed as a supplement to, or, in certain cases, a replacement for costly human MT evaluations [3][4][5][6]. These automatic evaluation methods rely on the idea that the quality of an MT can be measured by its similarity to a professional human translation. With each of the currently available automatic evaluation methods, this similarity is measured using a word-error metric between the sentences in the MT-produced text and the sentences in one or more human reference translations. The success of automatic MT evaluation depends largely on the amount of available comparable material and on the number of human reference translations, with more reference translations resulting in a more accurate measure of system performance.

In order to be able to use these metrics to calculate the similarity between a user query and a topic document in IR, we regard the original topic document and the MT-translated user query as translations of an unknown source text, as is shown in Figure 1.

The same three sets of topic translation were used as in the previous experiments. The topic translations and documents were pre-processed to remove stopwords, capitalisation and punctuation.

If we think of the query translations as human reference translations, it is possible to measure

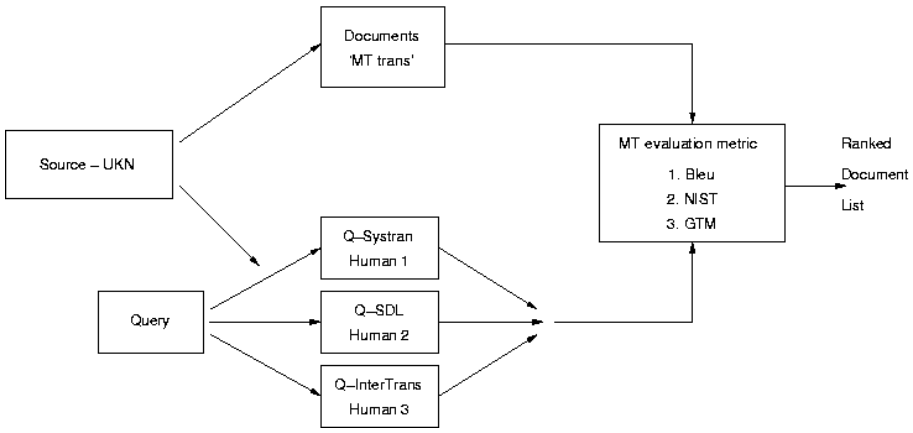

Figure 1: Document scoring based on MT Evaluation metrics.

the accuracy of the would-be 'machine translations' (the documents) using automatic MT evaluation metrics. The best 'machine translation' is the translation with the lowest word-error score with regard to the reference translations. The goal of our experiment was to find out to what extent the best 'machine translation' corresponded with a relevant document.

Experiments with development topics showed that best results were obtained with a combination of 2 existing MT evaluation methods (NIST and GTM) and an adaptation of the BLEU evaluation metric.

BLEU ranks different MT output texts based a combination of an N-gram similarity score and a sentence brevity penalty with respect to a corpus of human reference translations. The BLEU evaluation script was adapted in two ways. First, we eliminated the sentence brevity penalty. The original BLUE metric penalizes short sentences to avoid the possibility that very short segments such as 'the' would receive a maximum score when compared to any sentence containing 'the'. This penalty is clearly not relevant for the retrieval task at hand. A second modification to the script consisted in allowing a non-zero BLEU score, regardless of the fact that for one or more of the N-gram categories (unigram to 4-gram) no positive matches were found between the MT output and human reference translations.

NIST differs from BLEU with respect to both the co-occurrence score and the sentence brevity penalty. NIST alters the co-occurrence score in favour of lower order N-grams (i.e. low trigrams or quadrigram matches play less a role in the overall score) and more informative N-grams (i.e. N-grams that occur less frequently receive a higher weight). The sentence brevity penalty used by NIST is less severe than the one used by BLEU for sentences with small variations with respect to the reference translation.

GTM allows the calculation of standard precision and recall scores for automatically produced translations. It also calculates an f-measure score, which combines both the precision and recall scores for a given translation. It is this f-measure score, along with the NIST and adapted BLEU scores, that we used in our automatic ranking of the documents.

During our experiments we ranked the translated queries against the top 1000 documents retrieved for each topic using the PRF approach described in the previous section. We used a summation of the NIST, f-measure and adapted BLEU scores. We ran two sets of experiments. In the first set of experiments we evaluated the retrieved document list against only one reference translation, as produced by one of the three online MT systems, giving us three resulting ranking lists of documents for each topic. In a second set of experiments we merged the translated queries, using the three different translations of the topic as three different reference translations.

Table 4 shows results of document rescoring using MT evaluation metrics. Comparing these results to those using standard PRF methods in the earlier tables, it can be seen that the MT evaluation metrics are not effective for IR scoring in their present form. The main goal of our experiments was not to substantially improve the best available Image Retrieval methods, but

		SDL	INT	ST	MG
Dutch	Prec. 10 docs	0.116	0.172	0.124	0.140
	Av Precision	0.105	0.127	0.141	0.121
	Rel. Ret.	638	637	709	791
French	Prec. 10 docs	0.128	0.120	0.128	0.112
	Av Precision	0.107	0.110	0.117	0.100
	Rel. Ret.	666	707	658	695
German	Prec. 10 docs	0.164	0.172	0.124	0.148
	Av Precision	0.146	0.169	0.132	0.148
	Rel. Ret.	763	804	691	804
Italian	Prec. 10 docs	0.132	0.132	0.140	0.112
	Av Precision	0.132	0.119	0.118	0.108
	Rel. Ret.	633	591	602	639
Spanish	Prec. 10 docs	0.140	0.140	0.140	0.132
	Av Precision	0.145	0.111	0.128	0.131
	Rel. Ret.	767	666	649	755

Table 4: Retrieval runs with pseudo relevance feedback.

to investigate the novel idea for IR of treating topic documents and translated user queries as comparable translations of an unknown source text. Clearly based on the results shown here we need to explore further whether this approach can be adapted successfully for IR applications.

4 Conclusions and Further Work

Our experiments for ImageCLEF have demonstrated that our standard CLIR method works effectively for the short text documents in the St Andrew's collection, and further that there is potential for improvement in retrieval effectiveness from the use of image matching in CL image retrieval. Our experiments using MT evaluation metrics for scoring CLIR have so far not been successful, but we will be analysing our results to better understand the results and to seek alternative means of applying this approach.

References

[1] A. M. Lam-Adesina and G. J. F. Jones. Exeter at CLEF 2003: Experiments with Machine Translation for Monolingual, Bilingual and Multilingual Retrieval. In *Proceedings of Workshop of the Cross-Language Evaluation Forum (CLEF 2003)*, Trondheim, Norway, C. Peters et al. editors, Springer-Verlag, 2004.

[2] A. M. Lam-Adesina and G. J. F. Jones. Applying Summarization Techniques for Term Selection in Relevance Feedback. In *Proceedings of the 24th Annual International ACM SIGIR*, pages 1-9, New Orleans, ACM, 2001.

[3] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318, Philadelphia, USA, 2002.

[4] G. Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Human Language Technology: Notebook Proceedings: 128-132. San Diego, 2002.

[5] General Text Matcher `http://nlp.cs.nyu.edu/GTM/`

[6] NIST's MT Evaluation Toolkit `http://www.nist.gov/speech/tests/mt/resources/scoring.htm`