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

Gareth Jones, Declan Groves, Anna Khasin,

Adenike Lam-Adesina, Bart Mellebeek, Andy Way

School of Computing, Dublin City University, Ireland

Overview

- Aims of Participation
- Basic Retrieval Approach
- Standard Text Retrieval
- Text and Image Combination
- Automatic Machine Translation Metrics
- Conclusions

Aims of Participation

ImageCLEF St Andrew's task is interesting because:

- Documents and topics are short - high chance of term mismatch.
  - Rather more like early title and keyword retrieval tasks than current full-text retrieval.
- The likely importance of the image itself in determining document relevance.

Aims of Participation

Three sets of experiments which:

- Examine the effectiveness of our standard bilingual text retrieval system on this task.
- Make a preliminary investigation of the combination of text and image matching scores for this task.
- Explore the use of established automatic machine translation evaluation metrics in CLIR.

Basic Retrieval Approach

- Retrieval using the City University research distribution of the Okapi system.
  - Around 260 stop words removed from the texts, Porter stemming applied, small set of standard synonyms.
- Okapi augmented with summary-based pseudo relevance feedback (PRF) (Lam-Adesina & Jones SIGIR 2001).
- PRF adds 20 terms to original topic statement; original terms upweighted by a factor of 3.5.

Basic Retrieval Approach

Topics translated into English using three online machine translation resources:

- Systran (ST)
- FreeTranslation (SDL)
- InterTrans (INT)

Fourth translated topic statement formed by forming a union merge of the three translations (MG).

Standard Text Retrieval

		SDL	INT	ST	MG
Dutch	Av Precision Rel. Ret.	0.398 683	0.273 637	0.432 709	0.421 791
French	Av Precision Rel. Ret.	0.409 666	0.466 707	0.431 658	0.399 695
German	Av Precision Rel. Ret.	0.501 763	0.468 804	0.474 691	0.531 804
Italian	Av Precision Rel. Ret.	0.366 633	0.288 591	0.438 602	0.351 639
Spanish	Av Precision Rel. Ret.	0.444 767	0.318 666	0.406 649	0.398 755

Standard Text Retrieval

Observations:

- Considerable variation in average precision and number of relevant documents retrieved for different machine translation systems.
- Little direct correlation between average precision and number of relevant documents retrieved.
- Summary-based feedback works better than full-document feedback even for these short documents.

Text and Image Combination

- Simple experiment to merge results of text and image retrieval systems.
- Linear sum of text results from previous experiments and provided results from the VIPER image retrieval system.
- Merged list reordered and scored for retrieval effectiveness.

Text and Image Combination

			SDL	INT	ST	MG
French	Text Only	Av Precision	0.409	0.466	0.431	0.399
		Rel. Ret.	666	707	658	695
	Combined	Av Precision	0.407	0.466	0.428	0.399
		Rel. Ret.	666	707	658	695
Italian	Text Only	Av Precision	0.366	0.288	0.438	0.351
		Rel. Ret.	633	591	602	639
	Combined	Av Precision	0.369	0.289	0.437	0.351
		Rel. Ret.	633	591	602	639

Text and Image Combination

- Good news: the combination of the image matching scores with the text matching scores does not degrade retrieval.
  - in fact they sometimes improve results!
- Bad news: very little change in performance compared to standard text retrieval.

Further work will focus on carrying out feature analysis and scoring for image data.

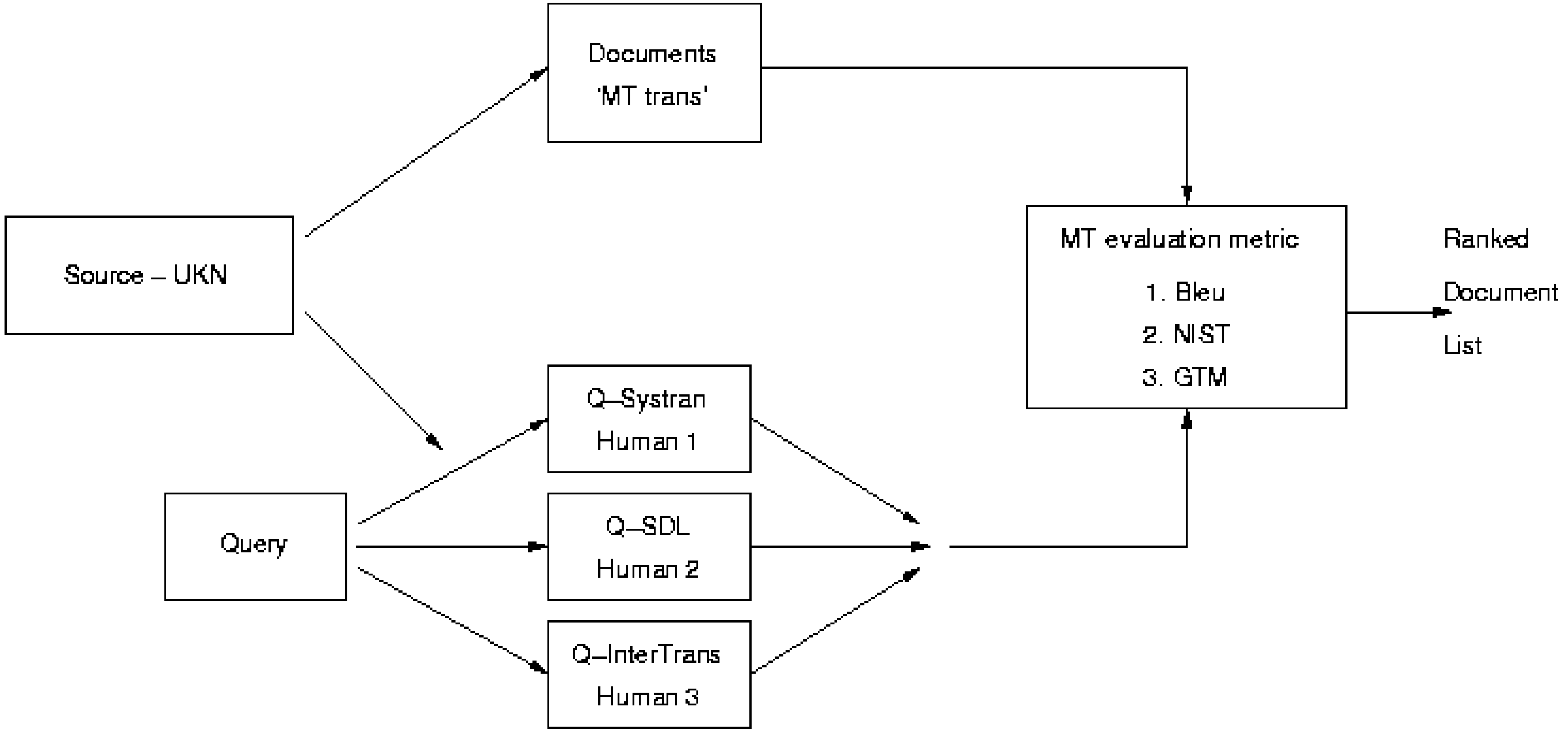
Automatic Machine Translation Metrics

- Automatic Machine Translation (MT) evaluation metrics are a supplement to costly human evaluation of MT systems.
- Based on the principle that the quality of an MT system can be measured by its similarity to a professional human translation.
- Current methods measure this similarity using a word-error rate metric between MT system output and one or more human reference translations.

Automatic Machine Translation Metrics

- The original document and the MT-translated user query are regarded as translations of an unknown source text.
- The translated topics are taken as human reference translations against which the accuracy of would-be MT output (the English documents) is calculated using MT evaluation metrics.
- The best MT is the one with the lowest word-error score with regard to the reference translation.
- Our goal of ImageCLEF experiments was to find out to what extent the best MT quality metrics correspond to document relevance.

Automatic Machine Translation Metrics



Document scoring based on MT Evaluation metrics.

Automatic Machine Translation Metrics

Three standard automatic machine translation metrics were investigated: BLEU, NIST, GTM.

- Top 1000 scoring documents from standard text retrieval system rescored using MT evaluation metric.
- The same three MT translated topics as in previous experiments.
- The documents and translated topics were processed to remove stopwords, capitalization, and punctuation.
- Various metrics tested on development set. Test runs using summation of BLEU, NIST and GTM.
- Separate runs on SDL, INT, ST and merged topic translations.

Automatic Machine Translation Metrics

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

Automatic Machine Translation Metrics

- The method used is currently much less effective than the standard text retrieval method.
- Further work is needed to explore whether MT evaluation metrics can be further adapted for effective complementary document scoring for CLIR.

Concluding Remarks

- Standard text CLIR methods are shown to be effective for the short documents in the St Andrew’s collection.
- There is potential to improve retrieval effectiveness by combining text retrieval with image matching, but further work is needed on this.
- MT evaluation metrics offer an alternative source of document to topic comparison information. At this stage we have not been able to utilize this information for effective CLIR.