Assessing Translation Quality for Cross Language Image Retrieval

Paul Clough and Mark Sanderson

Department of Information Studies, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK.
{p.d.clough,m.sanderson}@sheffield.ac.uk

Abstract. Like other cross language tasks, we show that the quality of the translation resource, among other factors, has an effect on retrieval performance. Using data from the ImageCLEF test collection, we investigate the relationship between translation quality and retrieval performance when using Systran, a machine translation (MT) system, as a translation resource. The quality of translation is assessed manually by comparing the original ImageCLEF topics with the output from Systran and rated by assessors based on their semantic content. Quality is also measured using an automatic score derived from the `mteval` MT evaluation tool, and compared to the manual assessment score. Like other MT tasks, we find that measures based on the automatic score are correlated with the manual assessments for this CLIR task. The results from this short study formed our entry to ImageCLEF 2003.

1 Introduction

Translating a user's search request from the *source language* of the query into the language of the document collection, the *target language*, is a core activity in Cross Language Information Retrieval (CLIR). Bridging the source-target translation gap can be achieved using a variety of translation resources, including bilingual dictionaries, extracting word/phrase equivalents from parallel or comparable corpora, machine translation (MT) or a controlled vocabulary. There are advantages and disadvantages to each approach, but commonly CLIR involves specialised knowledge of both CLIR and translation methodologies, and familiarity with the source and target languages.

As an information retrieval task, image retrieval involves translation to match user requests expressed in natural language to captions associated with the images which act as semantic representations of an image's visual content. As a CLIR task, image retrieval involves matching queries in the source language with captions in the target language. However, because the ImageCLEF test collection is new and previously unused for evaluation, we cannot be sure of the degree to which translation affects retrieval performance for the topics suggested in the proposed ad hoc retrieval task. As an image retrieval task there are other factors which affect whether retrieved images are relevant or not, such as the quality or size of the image, the quality of the caption

description, subjective interpretation of the image, and the short length of image descriptions.

For translation we use Systran, one of the oldest and most widely used commercial machine translation systems, freely available via a Web-based interface. Experience with this resource has shown that little or no multilingual processing is necessary as would normally be required when dealing with cross language retrieval, e.g. tokenisation, case and diacritic normalisation, decompounding and morphological analysis, therefore offering an attractive solution to problems involving translation. Systran has been used widely for CLIR before, including cross language image retrieval [2], but as a translation resource Systran presents limitations, such as one translation only for a source query and no control over translation.

In this paper we show how the quality of Systran varies across language and query, and illustrate some of the problems encountered when using Systran to translate the short ImageCLEF queries. These short texts of 2-3 words essentially use Systran for dictionary-lookup as they carry little grammatical structure to help translation. Although much previous research has already been undertaken in MT evaluation, there appears less empirical evaluation of translation quality within CLIR as translation quality is often judged based on retrieval performance. In this paper we measure translation quality as distinct from the retrieval.

The paper divides into the following: in section 2 we present background material, in section 3 the experimental setup, in section 4 the results, and in section 5 our conclusions and outline for future work.

2 Background

2.1 The ImageCLEF Task

ImageCLEF was a pilot experiment run at CLEF 2003 dealing with the retrieval of images by their captions in cases where the source and target languages differ (see [1] for further information about ImageCLEF). Because the document to be retrieved is both visual and textual, approaches to this task can involve the use of both multimodal and multilingual retrieval methods. The primary task at this year's ImageCLEF was an ad hoc retrieval task in which fifty topics were selected for retrieval and described using a topic title and narrative. Only the title was translated into Dutch, Italian, Spanish, French, German, Spanish and Chinese (by NTU), and therefore suitable for CLIR. As well as query-caption translation, further challenges for this task include: (1) captions which are typically short in length, (2) images that vary widely in their content and quality, and (3) short user search requests which provide little context for translation.

2.2 Systran

As a translation system, Systran is considered by many as a direct MT approach, although the stages resemble a transfer-based MT system because translation also involves the use of rules to direct syntax generation (see, e.g. [4]). There are essentially three stages to Systran: analysis, transfer and synthesis. The first stage, analysis, preprocesses the source text and performs functions such as character set conversion, spelling correction, sentence segmentation, tokenisation, and POS tagging. Also during the analysis phase, Systran performs partial analysis on sentences from the source language, capturing linguistic information such as predicate-argument relations, major syntactic relationships, identification of noun phrases and prepositional phrase attachment using their own linguistic formalism and dictionary lookup.

After analysis of the source language, the second process of transfer aims to match with the target language through dictionary lookup, and then apply rules to re-order the words according to the target language syntax, e.g. restructure propositions and expressions. The final synthesis stage cleans up the target text and determines grammatical choice to make the result coherent. This stage relies heavily on large tables of rules to make its decisions. For more information, consult [3] and [11].

2.3 MT Evaluation

Assessing how well an MT system works offers a challenging problem to researchers (see, e.g. [4] and [5]), and before evaluating an MT system, one must first determine its intended use and then evaluate the output based on whether the output is satisfactory for this purpose or not. MT evaluation is a subjective process and finding an objective measure is a non-trivial task. Dorr et al. [5] suggest that MT system evaluation can be treated similar to that of a software system where one evaluates the accuracy of input/output pairs (a *black-box* approach), or evaluates the data flow between internal system components (a *glass-box* approach).

In the black-box approach, a number of dimensions must be specified along which to evaluate translation quality (see, [5] for more information). In the glass-box approach, evaluation of system components might include linguistic coverage, or parsing accuracy. Organisations such as DARPA and NIST have established the necessary resources and framework in which to experiment with, and evaluate, MT systems as part of managed competitions, similar to the TREC (see, e.g. [12]) and CLEF (see, e.g. [9]) campaigns. For manual evaluation[1], three dimensions upon which to base judgments include translation *adequacy*, *fluency* and *informativeness*. Translation quality is normally assessed across an entire document when measuring fluency and informativeness, but adequacy is assessed between smaller units (e.g. paragraphs or sentences) which provide a tighter and more direct semantic relationship between bilingual document pairs. This is discussed further in section 3.1.

[1] See, e.g. TIDES: http://www.ldc.upenn.edu/Projects/TIDES/ [site visited: July 2003].

Test-suites can be used for both black-box and glass-box evaluation, and used to categorise the successes or failures of the system. The test-suite is often built for a specific application and type of evaluation in mind, and offers the research community a standardised resource within which different translation systems can be compared. Evaluation often takes the approach whereby the output of the MT system is captured and compared with a reference or gold-standard source and translation errors categorised and quantified, including lexical, grammatical and stylistic ones (see, e.g. [7]).

As well as manual methods of translation evaluation, there has also been much work in automating the task to reduce the amount of manual effort required, resulting in evaluation tools such as `mteval` which we discuss in section 3.2. The success of translation in CLIR is often based on retrieval performance and observations of translations, although previous work that does evaluate MT output as distinct from the retrieval process includes Patterson [8].

3 Experimental setup

3.1 Manual Assessment of Translation Quality

In these experiments, we have used the evaluation framework as provided by NIST for both manual and automatic evaluation. To assess adequacy, a high quality reference translation and the output from an MT system are divided into segments to evaluate how well the meaning is conveyed between versions. Fluency measures how well the translation conveys its content with regards to how the translation is presented and involves no comparison with the reference translation. Informativeness measures how well an assessor has understood the content of a translated document by asking them questions based on the translation and assessing the number answered correctly.

Given topic titles from the ImageCLEF test collection, we first passed them through the on-line version of Systran to translate them into English, the language of the image captions. We then asked assessors to judge the adequacy of the translation by assuming the English translation would be submitted to a retrieval system for an ad hoc task. Translators who had previously been involved with creating the ImageCLEF test collection were chosen to assess translation quality because of their familiarity with the topics and the collection, each assessor given topics in their native language. Translators were asked to assess topic titles[2] in the source language with the Systran English version and make a judgment on how well the translation captured the meaning of the original (i.e. how *adequate* the translated version would be for retrieval purposes). A five-point scale was used to assess translation quality, a score of 5 representing a very good translation (i.e. the same or semantically-equivalent words and

[2] In cases of multiple translations, we used the first.

[3] We used mteval-v09.pl which can be downloaded from: http://www.nist.gov/speech/tests/mt [site visited: July 2003]

syntax), to very bad (i.e. no translation, or the wrong words used altogether). Assessors were asked to take into account the "importance" of translation errors in the scoring, e.g. for retrieval purposes, mis-translated proper nouns might be considered worse than other parts-of-speech.

Table 1 shows an example topic title for each language and translation score for very good to good (5-4), okay (3) and bad to very bad (2-1) to provide an idea of the degree of error for these adequacy scores. We find that assessment varies according to each assessor; some being stricter than others, which suggests, further manual assessments may help to reduce subjectivity. In some cases, particularly Spanish, the source language title contains a spelling mistake which will affect translation quality. Some assessors allowed for this in their rating, others did not, therefore suggesting the need to manually check all topics for errors prior to evaluation.

Table 1 also highlights some of the errors produced by the MT system: (1) untranslated words, e.g. "*Muzikanten* and their instruments", (2) incorrect translation of proper nouns, e.g. "Bateaux sur Loch Lomond" translated as "Boats on Lomond *Log*" and "Il monte Ben Nevis" translated as "the mount *Very* Nevis", (3) mis-translations, e.g. "damage de guerre" translated as "*ramming* of war", and (4) wrong sense selection, e.g. "Scottish blowing chapels" where *kapelle* is mis-translated as chapel, rather than the correct word band. From this study, we found that many un-translated terms, however, were caused by mistakes in the original source texts. This might be seen as an additional IR challenge in which the queries reflect more realistic erroneous user requests. Systran was able to handle different entry formats for diacritics which play an important part in selecting the correct translation of a word, e.g. in the query "Casas de te' en la costa" (tearooms by the seaside), the word *te'* is translated correctly as *té* (sea) rather than *te* (you).

3.2 Automatic Assessment of Translation Quality

Although most accurate (and most subjective), manual evaluation is time-consuming and expensive, therefore automatic approaches to assess translation quality have also been proposed, such as the NIST `mteval`[3] tool. This approach divides documents into segments and computes co-occurrence statistics based on the overlap of word n-grams between a reference translation produced manually and an MT version. This method has been shown to correlate well with adequacy, fluency and informativeness because n-grams capture both lexical overlap and syntactic structure [4].

In the latest version of `mteval`, two metrics are used to compute translation quality: IBM's BLEU and NIST's own score. Both measures are based on n-gram co-occurrence, although a modified version of NIST's score has been shown to be the preferred measure [4]. These scores assume that the reference translation is of high quality, and that documents assessed are from the same genre. Both measures are influenced by changes in literal form. Translations with the same meaning but using different words score lower than those that appear exactly the same. This is justified in

assuming the manual reference translation is the "best" translation possible and the MT version should be as similar to this as possible.

Table 1. Example adequacy ratings assigned manually

Source	Adequacy rating	Original source	Systran English	Reference English
Chinese (simplified)	4-5	圣安德鲁斯风景的明信片	Saint Andrews scenery postcard	Picture postcard views of St Andrews
	3	战争造成的破坏	The war creates destruction	Damage due to war
	1-2	大亚茅斯海滩	Asian Mao si beach	Great Yarmouth beach
Dutch	4-5	Mannen en vrouwen die vis verwerken	men and women who process fish	men and women processing fish
	3	Vissers gefotografeerd door Adamson	Fisherman photographed Adamson	Fishermen by the photographer Adamson
	1-2	Muzikanten en hun instrumenten	Muzikanten and their instruments	Musicians and their instruments
German	4-5	Baby im Kinderwagen	Baby in the buggy	A baby in a pram
	3	Portät der schottischen Königin Mary	Portraet of the Scottish Queen Mary	Portraits of Mary Queen of Scots
	1-2	Museumaustellungsstücke	Museumaustellungs stuecke	Museum exhibits
French	4-5	La rue du Nord St Andrews	The street of North St Andrews	North Street St Andrews
	3	Bateaux sur Loch Lomond	Boats on Lomond log	Boats on Loch Lomond
	1-2	Damage de guerre	Ramming of war	Damage due to war
Italian	4-5	Banda Scozzese in Marcia	Scottish band in march	Scottish marching bands
	3	Vestito tradizionale gallese	Dressed traditional Welshman	Welsh national dress
	1-2	Il monte Ben Nevis	The mount Very Nevis	The mountain Ben Nevis
Spanish	4-5	El aforo de la iglesia	Chairs in a church	Seating inside a church
	3	Puentes en la carretera	Bridges in the highway	Road bridges
	1-2	las montañas de Ben Nevis	Mountains of Horseradish tree Nevis	The mountain Ben Nevis

For n-gram scoring, the NIST formula is:

$$Score = \sum_{n=1}^{N} \left\{ \frac{\sum_{\substack{all\ w_i \ldots w_n \\ that\ co\text{-}occur}} Info(w_1 \ldots w_n)}{\sum_{\substack{all\ w_i \ldots w_n \\ in\ sys\ output}} (1)} \right\} . \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\overline{L_{ref}}} \right) \right] \right\} \qquad \textbf{(1)}$$

where:

β is chosen to make the brevity penalty factor = 0.5 when the number of words in the system output is 2/3 of the average number of words in the reference translation.

N is the n-gram length.

$\overline{L_{ref}}$ is the average number of words in a reference translation, averaged over all reference translations.

L_{sys} is the number of words in the translation being scored.

$$Info(w_i \ldots w_n) = \log_2 \left(\frac{number\ of\ occurrences\ of\ w_1 \ldots w_{n-1}}{number\ of\ occurrences\ of\ w_1 \ldots w_n} \right)$$

The NIST formula uses $info(w_1 \ldots w_n)$ to weight the "importance" of n-grams based on their length, i.e. that longer n-grams are less likely than shorter ones, and reduces the effects of segment length on the translation score. The information weight is computed from n-gram counts across the set of reference translations. The brevity penalty factor is used to minimise the impact on the score of small variations in the length of a translation. The `mteval` tool enables control of the n-gram length and maximises matches by normalising case, keeping numerical information as single words, tokenising punctuation into separate words, and concatenating adjacent non-ASCII words into single words.

To evaluate the translation produced by Systran with `mteval`, we compared the English ImageCLEF topic title (the reference translation) with the English output from Systran (the test translation). Within the reference and test translation files, each topic title is categorised as a separate segment within a document, resulting in a NIST score for each topic. An alternative approach would be to treat the topics as separate segments within one document, although in practice we found the scores to be similar to those obtained from the first approach. To minimise the effects of syntactic variation on the NIST scores, we use an n-gram length of 1 word. For example, the English topic title "North Street St Andrews" is translated into French as "La rue du Nord St

Andrews" which translated literally into English is "The street of the North, St Andrews" which is rated as a good translation manually, but using an n-gram length > 1 would result in a low NIST score because of differences in word order.

3.2 The GLASS Retrieval System

At Sheffield, we have implemented our own version of a probabilistic retrieval system called GLASS, based on the "best match" BM25 weighting operator (see, e.g. [6]). Captions were indexed using all 8 fields, which include a title, description, photographer, location and set of manually assigned index categories, and the default settings of case normalisation, removal of stopwords and word stemming used by the retrieval system. To improve document ranking using BM25, we used an approach where documents containing *all* query terms were ranked higher than any other. The top 1000 images and captions returned for each topic title formed our entry to ImageCLEF. We evaluate retrieval effectiveness using *average precision* for each topic and across topics using *mean average precision* (or MAP) based on the ImageCLEF test collection.

4 Results and Discussion

4.1 Translation Quality

The average manual translation adequacy score across all languages for each topic is shown in Fig. 1. As one would expect, the average score varies across each topic, ranging from a minimum average score of 1.51 for topic 48 (Museum exhibits), to a maximum of 4.64 for topic 22 (Ruined castles in England), with an average manual score of 3.17 (i.e. okay).

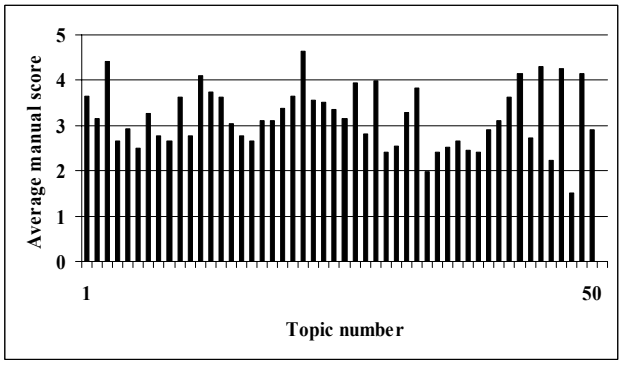

Fig. 1. Average manual adequacy score across the six languages for each topic

The six topics with an average score > 4 are topics 3 (Picture postcard views of St Andrews), 12 (North Street St Andrews), 22 (Ruined castles in England), 43 (British windmills), 45 Harvesting), 47 (People dancing) and 49 (Musicians and their instruments). The topics with an average score ≤ 2 are 34 (Dogs rounding-up sheep) and 48 (Museum exhibits). Example translations of topics 34, 48 and 22 are given in Table 2.

Table 2. Example translations of topics 34, 48 and 22

English:	*Dogs rounding up sheep*	*Museum exhibits*	*Ruined castles in England*
Italian	Dogs that assemble sheep	Exposures in museums	Ruins of castles in England
German:	Dogs with sheep hats	Museumaustellungssteucke	Castle ruins in England
Dutch:	Dogs which sheep bejeendrijven	Museumstukken	Ruin of castles in United Kingdom
French:	Dogs gathering of the preois	Exposure of objects in museum	Castles in ruins in England
Spanish:	Dogs urging on ewes	Objects of museum	Castles in ruins in England
Chinese:	Catches up with the sheep the dog	*no translation*	Become the ruins the English castle

Not only do the average translation adequacy scores vary across topics as shown in Fig. 1, the scores also vary across language as shown by the bar charts in Fig. 2. Although from Fig. 1 one can determine on average which topics perform better or worse, the results of Fig. 2 show that between languages results can vary dramatically (e.g. topic 2) based on at least three factors: (1) the translation resource, (2) the assessor's judgment for that topic, and (3) the difficulty of the topic itself to translate (e.g. whether it uses colloquial language, or expresses a more general or specific concept). Some topics, such as topic 22 (Ruined castles in England) score similarly between all languages, but in general we observe that translation quality varies across topic and language (see also Table 1).

Table 3 summarises translation quality for both the manual and automatic assessment. On average Spanish, German and Italian translations are rated the highest manually indicating these are the strongest to-English Systran bilingual pairings; Chinese, Dutch and French are the lowest suggesting the weakest pairs. The Systran translations for Chinese are on average the shortest and 14% of topics get a rating very bad (the third highest), and 28% of topics a rating of very good (the lowest). French has the highest number of topics rated very poor, followed by Chinese and Italian. Upon inspection, many of these low scores are from words which have not been translated.

The bar chart in Fig. 3 shows the average NIST score across all languages for each topic, and immediately we see a much larger degree of variation across topics than for the manual scores shown in Fig. 1.

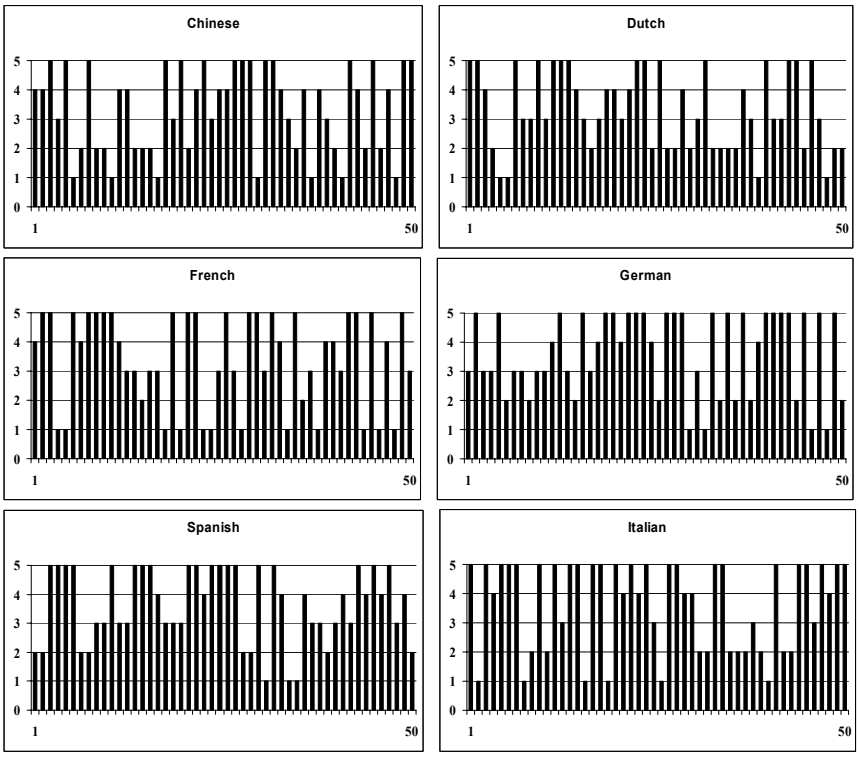

Fig. 2. Manual translation adequacy scores for each language and topic

Table 3. A summary of manual and automatic topic assessment for each source language

	Avg man score	Avg NIST score	man-NIST correla-tion	Mean translation length (words)	% topics man = 1	% topics man = 5	% topics NIST = 0
Chinese	3.34	1.68	0.268*	3.76	14%	28%	38%
Dutch	3.32	3.27	0.426*	4.32	8%	30%	12%
German	3.64	3.67	0.492*	3.96	8%	44%	10%
French	3.38	3.67	0.647*	4.78	24%	40%	8%
Italian	3.65	2.87	0.184	5.12	12%	50%	18%
Spanish	3.64	3.24	0.295*	4.38	6%	34%	10%

*Spearman's rho correlation significant at $p<0.01$

Overall, the highest automatic scores which are ≥ 5 are achieved with topics 1 (men and women processing fish), 23 (London bridge), 26 (Portraits of Robert Burns) and 49 (Musicians and their musical instruments). Topics with scores ≤ 1 are 5 (woodland

scenes), 46 (Welsh national dress) and 48 (museum exhibits). Low scores are often the result of variation in the ways in which concepts are expressed in different languages. For example, in Chinese the query "coat of arms" is interpreted as "a shield" or "heraldic crest" because a direct equivalent to the original English concept does not exist. When translated back to English using Systran, more often than not the query is translated literally resulting in a low word overlap score.

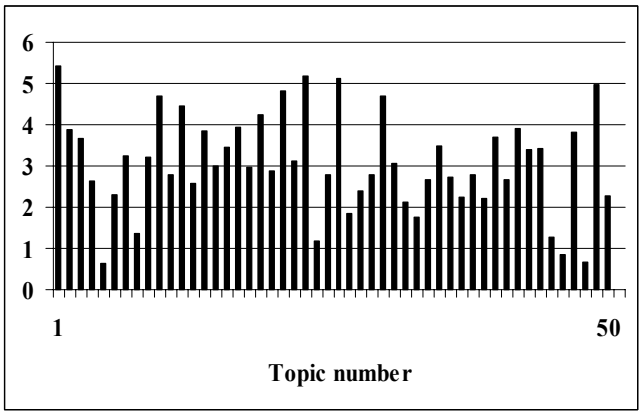

Fig. 3. Average automatic NIST adequacy score across all languages for each topic

From Table 3, Chinese also has the lowest average NIST score (1.68), which can be explained by the large proportion of topics with a zero score (38%) and the shorter average query length. Of these 38% of topics with a score of 0, 37% have no translation from Systran. From Table 3, German and French have the highest average NIST score, followed by Dutch and Spanish.

Contributing to the low Spanish scores is the high number of spelling errors in the source queries which result in non-translated words. Table 4 shows example translations with a 0 NIST score, i.e. where the reference and Systran translations have no words which overlap. In many cases, however, this is simply because different words are used to express the same concept, or lexical variations of the word (such as plurals) are used instead. For information retrieval, this is important because if a simple word co-occurrence model is used with no lexical expansion; the queries may not match documents (although in some cases stemming would help). This highlights the limitation of using `mteval` for assessing translation quality in CLIR because comparison is based on literal word overlap only.

Table 4. Example translations with a NIST score of 0

	Reference translation	Systran version	Man score
Chinese	Woodland scenes	Forest scenery	5
	Scottish marching bands	*no translation*	1
	Tea rooms by the seaside	Seashore teahouse	5
	Portraits of Mary Queen of Scots	*no translation*	1
	Boats on Loch Lomond	In Luo river Mongolia lake ships	2
	Culross abbey	Karohs overhaul Daoist temple	3
	Road bridges	Highway bridge	5
	Ruined castles in England	Becomes the ruins the English castle	4
	Portraits of Robert Burns	*no translation*	4
	Male portraits	Men's portrait	5
	The mountain Ben Nevis	Nepali Uygur peak	2
	Churches with tall spires	Has the high apex the churches	4
	A coat of arms	*no translation*	1
	British windmills	England's windmill	4
	Waterfalls in Wales	Well's waterfall	2
	Harvesting	Harvests	5
French	Woodland scenes	Scenes of forests	1
	Waterfalls in Wales	Water falls to the country of Scales	1
	Harvesting	Harvest	5
	Mountain scenery	Panorama mountaineer	3
German	Glasgow before 1920	*no translation*	1
	Male portraits	Portraets of men	1
	Harvesting	Harvests	5
	Welsh national dress	Walisi tract	1
	Museum exhibits	Museumaustellungsstuecke	1
Italian	Woodland scenes	Scene of a forest	5
	Tea rooms by the seaside	It knows it from te' on lungo-mare	1
	Wartime aviation	Air in time of war	4
	British windmills	English flour mills	2
	Welsh national dress	Dressed traditional Welshman	3
	People dancing	Persons who dance	5
Spanish	Woodland scenes	A forest	5
	Wartime aviation	Aviators in time military	2
	Male portraits	Picture of a man	4
	Museum exhibits	Objects of museum	2
	Mountain scenery	Vista of mountains	1
Dutch	Woodland scenes	bunch faces	1
	Road bridges	Viaducts	4
	Men cutting peat	Turfstekers	1
	Mountain scenery	Mount landscapes	2

These differences also contribute to the lack of statistical correlation for topics between the manual and automatic assessments (shown in Table 3). Using Spearman's rho to indicate in general whether the same topics are assigned a high or low score for both manual and automatic assessments at a statistical significance of $p<0.01$, we find that Chinese and Spanish have lowest significant correlation. For Chinese this is caused by the high number of topics with no translation, and Spanish because of spelling errors resulting in non-translated terms.

The correlation between scores for Italian is not significant which upon inspection is found to the use of different words from the original English to describe equivalent translations. Another contributing factor is the query length, which is generally longer (see Table 3) because of a more descriptive nature, e.g. "men cutting peat" (English) is translated as "men who cut the peat" (Italian). A further cause of non-correlation comes from words which are not translated, e.g. "Portraits of Robert Burns" (English) and "Ritratto of Robert Burns". Topics containing non-translated words are given a low manual score, but in the previous example 3 of the 4 original English terms are present which gives a high NIST score. For Dutch topics, erroneous translations are also caused by the incorrect translation of compounds (which also occurs in German). For example, the German compound "eisenbahnunglueck" is not translated.

4.2 Retrieval Performance

Fig. 4 shows a graph of recall versus precision across all topics and for each language using the *strict intersection* set of ImageCLEF relevance judgments. As with previous results for CLIR tasks, monolingual performance is the highest. Chinese has the lowest precision-recall curve, and is noticeably lower than the rest of the languages which group together and follow a similar shape. The French curve is the highest of the languages, which matches with Table 3 where French has the lowest NIST score, the least number of topics with a zero NIST score, and a high proportion of topics with a high manual assessment rating.

Fig. 5 provides a breakdown of retrieval performance across topics. The stacked bar chart shows monolingual average precision, and average precision across cross language results for each topic. Some languages perform better or worse for each topic (depending on the quality of translation), but the graph provides an overall indication of those topics which perform better or worse. Across all languages (excluding English) and topics, the MAP score is 0.420 (with a standard deviation of 0.23) which is on average 75% of monolingual performance (Table 5 shows the breakdown across languages).

Topics which perform poorly include 4 (seating inside a church), 5 (woodland scenes), 29 (wartime aviation), 41 (a coat of arms) and 48 (museum exhibits). These exhibit average NIST scores of 2.63, 0.64, 2.80, 3.71 and 3.83 respectively, and manual ratings of 3, 3.7, 4.17, 3.5 and 1.83 respectively. In some cases, the translation quality is high, but the retrieval low, e.g. topic 29, because relevance assessment for cross language image retrieval is based upon the image and caption. There are cases

when images are not relevant, even though they contain query terms in the caption, e.g. the image is too small, too dark, the object of interest is obscured or in the background, or the caption contains words which do not describe the image contents (e.g. matches on fields such as the photographer, or notes which provide background meta-information).

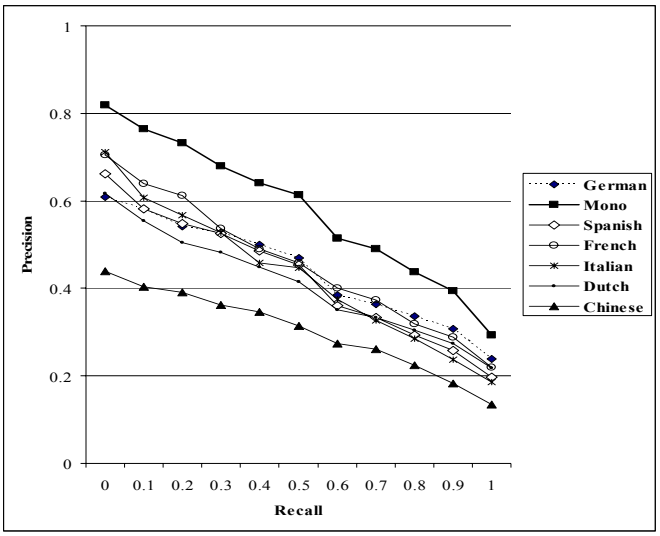

Fig. 4. Precision-recall graph for the Sheffield ImageCLEF entry

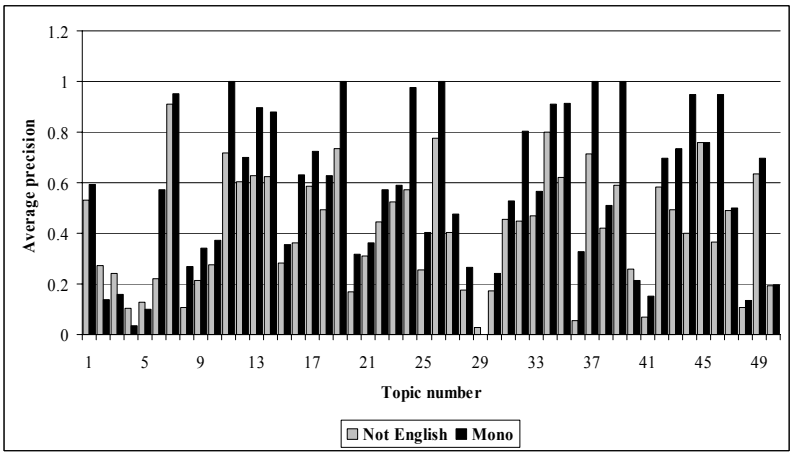

Fig. 5. Monolingual average precision and MAP across systems (excluding English)

Topic 29 (wartime aviation) and 4 (seating in a church) have very low monolingual average precision scores. For topic 29 this is because relevant images do not contain the terms "wartime" or "aviation", but rather terms such as "war", "planes", "runway" and "pilot". Relevant images for topic 29 relied on manual assessors using the interactive search and judge facility. We also find that the differences between the language of the collection and translated queries contribute to low average precision scores. This comes from two sources: (1) manual query translation and (2) the dictionary used by Systran. For example in Italian, the query "windmill" is translated manually as "mill of wind" which would match "wind" and "mill" separately. However, most captions only contain the term "windmill" and therefore do not match a query containing "wind" and "mill". The query "road bridge" is translated by Systran as "highway bridge" which will not match the collection because the only reference to a highway refers to a footpath and not a road.

Table 5. A summary of retrieval performance and possible influences on retrieval

	MAP	%mono	Avg Prec - man	Avg Prec - NIST	Avg Prec - query len	Avg Prec - #relevant
Chinese	0.285	51%	0.472*	0.384*	0.370*	0.159
Dutch	0.390	69%	0.412*	0.426*	0.374*	-0.165
German	0.423	75%	0.503*	0.324*	0.133	-0.281
French	0.438	78%	0.460*	0.456*	0.022	-0.046
Italian	0.405	72%	0.394*	0.378*	-0.011	-0.098
Spanish	0.408	73%	-0.061	0.462*	-0.025	0.025
Mono	0.562	-	-	-	-	-

*Spearman's rho correlation significant at $p<0.01$

Table 5 summarises retrieval performance for each language and Spearman's rho between average precision and a number of possible influences on retrieval for each topic. We find that French has the highest MAP score (78% monolingual), followed by German (75% monolingual) and Spanish (73% monolingual). In general, average precision and translation quality is correlated (using Spearman's rho with $p<0.01$) for both the manual and automatic assessments which suggests that a higher quality of translation does give better retrieval performance, particularly for Chinese, German and French (manual assessments) and Spanish, French and Dutch (automatic assessments). The correlation between the manual scores and average precision scores is not significant and we find this is because of spelling errors in the Spanish source texts. In general the length of query and number of relevant document for a topic does not affect retrieval, although query length does obtain significant correlation for Chinese and Dutch. This corresponds with these languages generally having longer and more varied translation lengths (Table 3).

We might expect the average precision scores to correlate well with the NIST score for the GLASS system because both are based on word co-occurrences, but it is interesting to note that retrieval effectiveness is correlated just as highly with the manual assessments (except Spanish), even though correlation between the manual and automatic assessments is not always itself high. This is useful as it shows that as a CLIR

task, the quality of translation in the ImageCLEF cross language image retrieval task has a significant impact on retrieval thereby enabling, in general, retrieval effectiveness to indicate the quality of translation.

5 Conclusions and Future Work

We have shown that cross language image retrieval for the ImageCLEF ad hoc task is possible with little or no knowledge of CLIR and linguistic resources. Using Systran requires little effort, but at the price of having no control over translation or being able to recover when translation goes wrong. In particular, Systran provides only one translation version which is not always correct and provides only one alternative. There are many cases when proper names are mistranslated, words with diacritics not interpreted properly, words translated incorrectly because of the limited degree of context and words not translated at all.

We evaluated the quality of translation using both manual assessments, and an automatic tool used extensively in MT evaluation. We find that quality varies between different languages for Systran based on both the manual and automatic score which is correlated, sometimes highly, for all languages. There are limitations, however, with the automatic tool which would improve correlation for query quality in CLIR evaluation, such as resolving literal equivalents for semantically similar terms, reducing words to their stems, removing function words, and maybe using a different weighting scheme for query terms (e.g. weight proper names highly). We aim to experiment further with semantic equivalents using Wordnet and collection-based equivalents, and also assess whether correlation between the manual and automatic scores can be improved by using longer n-gram lengths.

Using GLASS we achieve cross language retrieval at 75% of the monolingual average precision score. Although Chinese retrieval is lowest at 51%, this would still provide multi-lingual access to the ImageCLEF test collection, albeit needing improvement. As the simplest approach possible, the challenge for ImageCLEF is what can be done to improve retrieval above the baseline set by Systran. Given that the task is not purely text, but also involves images, retrieval could be improved using content-based methods of retrieval, post-translation query expansion based on relevance feedback, and pre-translation query expansion based on EuroWordnet, a European version of Wordnet, and the ImageCLEF collection.

As a retrieval task, we have shown that translation quality does affect retrieval performance because of the correlation between manual assessments and retrieval performance, implying that in general, higher translation quality results in higher retrieval performance. We have also shown that for some languages, the manual assessments correlate well with the automatic assessment suggesting this automatic tool could be used to measure translation quality given a CLIR test collection.

6 Acknowledgments

We would like to thank members of the NLP group and Department of Information Studies for their time and effort in producing manual assessments. Thanks also to Hideo Joho for help and support with the GLASS system. This work was carried out within the Eurovision project at Sheffield University, funded by the EPSRC (Eurovision: GR/R56778/01).

References

1. Clough, P. and Sanderson, M.: The CLEF 2003 cross language image retrieval task. This volume. (2003)
2. Flank, S.: Cross-Language Multimedia Information Retrieval. In Proceedings of Applied Natural Language Processing and the North American Chapter of the Association for Computational Linguistics (ANLP-NAACL2000). (2000)
3. Heisoft: How does Systran work? http://www.heisoft.de/volltext/systran/dok2/howorke.htm (2002)
4. Hutchins, W.J. and Somers, H.: An Introduction to Machine Translation. Academic Press, London, England (1986)
5. Jordan, P.W., Dorr, B.J. and Benoit, J.W.: A First-Pass Approach for Evaluating Machine Translation Systems. Machine Translation, **Vol. 8(1-2)** (1993) 49-58
6. National Institute of Standards and Technology (NIST): Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. http://www.nist.gov/speech/tests/mt/resources/scoring.htm (2002)
7. Nyberg, M. and Carbonell, J.: Evaluation Metrics for Knowledge-Based Machine Translation. In Proceedings of Fifteenth International Conference on Computational Linguistics (COLING-94). (1994)
8. Patterson, C.: The Effectiveness of Using Machine Translators to Translate German and Portuguese into English when Searching for Images with English Captions. MSc dissertation for the degree of Masters of Arts in Librarianship, Department of Information Studies, University of Sheffield. (2002)
9. Peters, C. and Braschler, M.: Cross-Language System Evaluation: The CLEF Campaigns. In Journal of the American Society for Information Science and Technology **Vol. 52(12)** (2001) 1067-1072
10. Robertson, S., Walker, S. and Beaulieu, M.: Okapi at TREC-7: automatic ad hoc, filtering VLC and interactive track. In NIST Special Publication 500-242: Proceedings of TREC-7. Gaithersburg, MA (1998) 253-264
11. Systran Ltd: The SYSRAN linguistics platform: A software solution to manage multilingual corporate knowledge. http://www.systransoft.com/Technology/SLP.pdf (2002)
12. Voorhees, E.M. and Harman, D.: Overview of TREC 2001, In NIST Special Publication 500-250: Proceedings of TREC2001, NIST. (2001)