DCU and UTA at ImageCLEFPhoto 2007

Anni Järvelin[1], Peter Wilkins[2], Tomasz Adamek[2], Eija Airio[1], Gareth J.F. Jones[2], Alan F. Smeaton[2] and Eero Sormunen[1]

[1] University of Tampere (UTA), Finland
Anni.Jarvelin@uta.fi
[2] Dublin City University (DCU), Ireland
pwilkins@computing.dcu.ie

Abstract. Dublin City University (DCU) and University of Tampere (UTA) participated in ImageCLEF 2007 photographic retrieval task with several monolingual and bilingual runs. The approach was language independent with text retrieval utilizing fuzzy s-gram query translation and combined with visual retrieval. Data fusion was achieved through unsupervised query-time weight generation approaches. The baseline was a combination of dictionary-based query translation and visual retrieval, which achieved the best result. The best mixed modality runs using fuzzy s-gram translation reached on average around 83% of the baselines' performance. This approach was much closer at the early precision levels of P@10 and P@20. This suggests that our language independent approach could be a cheap alternative for cross-lingual image retrieval. Both sets of results further emphasize the merit in our query-time weight generation schemes for data fusion, with the fused runs exhibiting marked performance increases over single modalities without the use of prior training data.

1 Introduction

Retrieving images by their associated text is a common approach in image retrieval [1]. When cross-language image retrieval is considered, this approach requires language dependent linguistic resources for query translation. Machine-readable dictionaries, machine translation tools or corpus-based translation tools are expensive and not available for all the language pairs. However, there are alternative approaches which may be used to compensate linguistic tools, for example the fuzzy string matching technique n-gram matching and its generalization s-gram matching. These techniques have previously been used for translation of query words missing from dictionaries [2][3], but only rarely for the whole query translation [4][5].

In earlier ImageCLEF campaigns combined text and visual retrieval approaches have performed better than text or image retrieval alone. In this year's campaign, text retrieval faced a new challenge of retrieval of lightly annotated photographs [1]. A negative impact on the performance of the text retrieval techniques was to be expected and therefore successful fusion of text and visual

features became even more important. We tested a language independent image retrieval approach, where s-gram-based fuzzy query translations were fused with visual retrieval. We explored data fusion techniques with query-time coefficient generation for retrieval expert combination. We experimented primarily with altering the stages at which we fuse various experts together. For instance we experimented with fusing all the visual experts into a single expert, then fusing with text, as opposed to treating all experts equally. To have a strong baseline, the performance of the language independent approach was compared to a combination of dictionary-based query translation and visual retrieval.

To study the effect of the source and target languages on the quality of the fuzzy translation we selected six language pairs where source/target languages were related to each other at different levels. The Scandinavian languages Danish, Norwegian and Swedish were translated into German, to which they are quite closely related to. French was the source language that was closest to English. German and English are not very closely related and translation between them was done both ways. (See [4] for a matrix for similarities between English, French, German and Swedish) A total of 138 runs were submitted. Reporting the results for all of these would be impractical and therefore only the results for the most interesting runs are presented here.

2 Background

s-gram-based query translation. The s-gram matching is a fuzzy string matching technique that measures similarity between text strings. The text strings to be compared are decomposed into substrings (s-grams) and the similarity is calculated as the overlap of their common substrings. s-gram matching is a generalization of n-gram technique, commonly used for matching cognates in Cross Language Information Retrieval (CLIR). While the n-gram substrings consist of adjacent characters of the original strings, skipping some characters is allowed when forming the s-grams. In classified s-gram matching [6], different types of s-grams are formed by skipping different number of characters. The s-grams are then classified into sets based on the number of characters skipped and only the s-grams belonging to the same set are compared to each other when calculating the similarity. Character Combination Index (CCI) indicates the set of all the s-gram types to be formed from a string. CCI $\{\{0\},\{1,2\}\}$ for example means that three types of s-grams are formed and classified into two sets: one set of conventional n-grams formed of adjacent characters ($\{0\}$) and one of s-grams formed both by skipping one and two characters ($\{1,2\}$). For an extensive description of the s-gram matching technique, see [6][7].

Proper names are very common query terms when searching from image databases [8] and are typically not covered by translation dictionaries and thus remain untranslatable in queries. Proper names in related languages are often spelling variants of each other and can thus be translated using approximate string matching.

Visual Retrieval. To facilitate visual retrieval we made use of six 'low-level'

global visual experts. Our visual features are MPEG7 features and were extracted using the aceToolBox, developed as part DCU's collaboration in the aceMedia project [9]. These six features included: Colour Layout, Colour Structure, Colour Moments, Scalable Colour, Edge Histogram and Homogenous Texture. Further details on these descriptors can be found in [10] and [11]. Distance metrics for each of these features were implementations of those specified in the MPEG7 specification [11].

Query-Time Fusion. The combination of retrieval experts for a given information need can be expressed as a *data fusion* problem [12]. Given that for any given information need different retrieval experts perform differently, we require some form of weighting scheme in order to combine experts. Typical approaches to weight generation include the use of global query-independent weights or query-class expert weights learnt through experimentation on a training collection to name a few.

Our approach to weight generation differs in that it is a query-dependant weighting scheme for expert combination which requires no training data [13]. This work was based upon an observation, that if we were to plot the normalized scores of an expert, against that of scores of other experts used for a particular query, that the expert who's scores exhibited the greatest initial change in scores correlated with that expert being the best performer for that query. Examples of this observation can be seen in [13] and our ImageCLEF workshop paper [14]. This approach is not giving us any absolute indication of expert performance, which other approaches to examining score distributions attempt to provide, an excellent review of which is given by Robertson [15]. We would note that this observation is not universal, and we can identify failure cases where this observation will not occur. If we assume though that in a majority of queries this observation will hold, then we can employ techniques that leverage this approach to create query-time expert coefficients for data fusion. Our main technique involves measuring the change in scores for a retrieval expert over a top subset of its results, versus the change in scores over a larger sample of that experts scores. The expert which undergoes the greatest initial change in score is assigned a greater weight. A complete explanation of this process can be found in [13].

3 Resources, Methods and Runs

Text Retrieval and Indexing. We utilized the Lemur toolkit (Indri engine) [16] for indexing and retrieval. Indri combines language modeling to inference nets and allows structured queries to be evaluated using language modeling estimates. The word tokenization rules used in indexing included converting punctuation marks into spaces and capitals were converted into lower case. Strings separated by the space character were tokenized into individual words. For the dictionary-based translation, lemmatized English and German indices were created. The image annotation text was lemmatized, words not recognized by the lemmatizers were indexed as such. Compound words were split and both the original compound and the constituents were indexed. For the s-gram-based

translation we used inflected indices, where the words were stored in the inflected word forms in which they appeared in the image annotations. Stop words were removed.

Topics were processed as the indices. For the s-gram-based translation stop words were removed from the queries and the remaining words were translated into the target language with s-gram matching. The CCI was set to be 0,1,2 and the Jaccard coefficient [7] was used for calculating the s-gram proximity. Three best matching index words were selected as translations for each topic word. A similarity threshold value of 0.3 was set to discard bad translations, only the index words exceeding this threshold with respect to a source word were accepted as translations. As a consequence some of the query words could not be translated. Queries were structured utilizing a synonym operator where target words derived from the same source word were grouped into the same synonym group (*the Pirkola method*, [17]). Indris Pseudo Relevance Feedback (PRF) was used with 10 keys from the 10 highest ranked documents in the original result list.

For the dictionary-based query translation, the UTACLIR query translation tool was used. UTACLIR was originally developed for the CLEF 2000 and 2001 campaigns [2]. It utilizes external language resources, such as translation dictionaries, stemmers and lemmatizers. Topic words were lemmatized, stop words removed and finally the non-stop words translated. Next, untranslatable compound words were split and the constituents were translated. Translation equivalents were normalized utilizing a target language lemmatizer. Untranslatable words were matched against the database index using the s-gram matching. Queries were structured with the synonym operator. A morphological analyzer for French was not available and therefore the French topics were analyzed manually. This might have resulted in a slightly better quality of lemmatization than automatic analysis, even though we strived for comparable quality. We used PRF with 20 expansion keys from the 15 top ranked documents.

Data Fusion. The query-time data fusion approach specified in Section 2 describes our basic approach to expert combination. However, one set of parameters that was not specified was the order in which experts will be combined. This is the focus of our experimental work in this section.

One commonality between all the combination approaches we try in this work is the fusion of the low-level visual experts. For each query image we fuse the six low-level visual experts into a single result for each image, where the combination of these is using the aforementioned technique. Therefore for each query, the visual component was then represented by three result sets, one for each query image. Additionally for a subset of our runs we introduce a seventh visual expert, the FIRE baseline [18]. In cases where FIRE was used, because it was a single result for the three visual query images, we first combined our MPEG7 visual features into a single result for each image, then these combined into a overall image result, which was then combined with the FIRE baseline. There are four variants that we tried in our combination work, which are:

– dyn-equal: Query-time weighting method used, text and individual image results combined at the same level (i.e. we have three image results and one text result which is to be combined).
– dyn-top: As above, except the results for each query image were fused into a single image result, which was then combined with the text result (i.e. image results combined into a single result, which is then combined with the single text result).
– stat-eventop: Query-time weighting to produce single image result list, image and text fused together with equal static weighting (0.5 coefficient).
– stat-imgHigh: As above, except with the image result assigned a static weight of 0.8 and text a static weight of 0.2.

Additionally, any of our runs which ended in 'fire' incorporated the FIRE baseline into the set of visual experts used for combination.

4 Results

Our tables of results are organized as follows. Table 1 presents our baseline runs, including monolingual text-only, visual-only and baseline fusion results mixing these two types. Table 2 presents our central cross-lingual results with mixed modalities. In all tables where data fusion is utilized, we present only the best performing data fusion approach. Except where noted, all visual results used in data fusion presented here incorporated the FIRE baseline as visual data which included the FIRE baseline with our global MPEG7 features consistently outperformed global MPEG7 by themselves.

Table 1. ImageCLEF Baseline Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
EN-EN	Text	dict	n/a	no	0.1305	0.1550	0.1408
EN-EN	Text	*s*-gram	n/a	yes	0.1245	0.1133	0.1242
DE-DE	Text	dict	n/a	yes	0.1269	0.1717	0.1533
DE-DE	Text	*s*-gram	n/a	yes	0.1067	0.1233	0.1125
MPEG7 With FIRE	Visual	na	dyn-equal	no	0.1340	0.3600	0.2658
MPEG7 Without FIRE	Visual	na	dyn-equal	no	0.1000	0.2700	0.1958
EN-EN	Mixed	dict	dyn-equal	yes	0.1951	0.3967	0.3150
EN-EN	Mixed	*s*-gram	dyn-equal	yes	0.1833	0.3833	0.3092
DE-DE	Mixed	dict	dyn-equal	yes	0.1940	0.4033	0.3300
DE-DE	Mixed	*s*-gram	dyn-equal	yes	0.1628	0.3350	0.2792

For the monolingual runs in Table 1, the runs where morphological analysis (dict) was used performed slightly better than the *s*-gram runs. The difference is small for the English runs. For German runs the difference is greater, which

is understandable as German has a more complex inflectional morphology than English. Our text and visual retrieval techniques were almost equal, which is notable in the context of earlier years' ImageCLEF results. Our best visual-only run performed well being the second best visual approach in terms of Mean Average Precision (MAP). Its MAP value 0.1340 is comparable to our best monolingual English text run scoring 0.1305. We believe that the comparative low performance of the text expert (when compared to the dominance of text in previous years) was due to the reduced length of the annotations for 2007. Table 1 also presents our fused monolingual text and visual retrieval runs, which performed clearly better than any of the text or visual runs alone. Fusion of these modalities produced consistent improvements in MAP of between 65% and 67%.

From a data fusion perspective, no single approach of the four we tried consistently performed the best. Whilst our results presented here show the "dyn-equal" fusion as being superior, this is because it was the only fusion type attempted with visual data which incorporated the FIRE baseline. For runs where FIRE was not used, there best performing fusion type varied depending on the text type (dictionary or sgram) or language pair used. In a majority of cases all fusion types performed similarly, as such deeper investigation with significance testing will be required in order to infer any meaningful interpretations. However, we can conclude that as all four fusion types made use of our query-time weight generation method at some level, that this technique is capable of producing weights which lead to performance improvements when combining results. What is unknown is how far from the optimal query-dependant combination performance we achieved, and that will be one of the ultimate measures of the success of this approach.

Table 2 presents our central cross-lingual results. Dictionary-based query translation was the best query translation approach. The best mixed modality runs using the *s*-gram-based query translation nevertheless reached on average around 84% of the MAP of the best mixed modality runs using dictionary-based translation. The difference between the approaches further decreased when the early precision values of P@10 and P@20 were considered. The best *s*-gram runs reached on average around 91% of the best dictionary-based runs performance at P@10 and around 89% at P@20. If the high ranks of the result list are considered to be important from the user perspective, the *s*-gram translation could be seen as almost equal with the dictionary-based translation in mixed modality runs. These results varied depending on the language pair. *s*-gram-based and dictionary-based translation performed similarly for the closely related language pairs, while the differences were greater for the more distant language pairs. The *s*-gram translation reached its best results with Norwegian and Danish topics and German annotations - over 90% of the dictionary translation's MAP, and the worst ones between German and English - less than 80% of the dictionary translation's MAP.

Table 2. ImageCLEF CLIR Fusion Results

Language Pair	Modality	Text	Fusion	FB	MAP	P@10	P@20
FR-EN	Mixed	dict	dyn-equal	yes	0.1819	0.3583	0.2967
FR-EN	Mixed	*s*-gram	dyn-equal	no	0.1468	0.3483	0.2667
DE-EN	Mixed	dict	dyn-equal	yes	0.1910	0.3483	0.3042
DE-EN	Mixed	*s*-gram	dyn-equal	yes	0.1468	0.3233	0.2533
DA-DE	Mixed	dict	dyn-equal	yes	0.1730	0.3467	0.2942
DA-DE	Mixed	*s*-gram	dyn-equal	yes	0.1572	0.3350	0.2717
NO-DE	Mixed	dict	dyn-equal	yes	0.1667	0.3517	0.2700
NO-DE	Mixed	*s*-gram	dyn-equal	yes	0.1536	0.3167	0.2667
SV-DE	Mixed	dict	dyn-equal	yes	0.1788	0.3817	0.2942
SV-DE	Mixed	*s*-gram	dyn-equal	yes	0.1472	0.3050	0.2500
EN-DE	Mixed	dict	dyn-equal	yes	0.1828	0.3633	0.3008
EN-DE	Mixed	*s*-gram	dyn-equal	yes	0.1446	0.3350	0.2667

5 Conclusions

In this paper we have presented the joint DCU and UTA ImageCLEF 2007 Photo results. In our work we experimented with two major variables, that of cross-lingual text retrieval utilizing minimal translation resources, and query-time weight generation for expert combination. Our results are encouraging and support further investigation into both approaches. Further work is now required to conduct a more thorough analysis of contributing factors to performance emphasized by each approach. Of particular interest will be the degree to which each of these approaches introduced new information, or re-ordered existing information presented by the systems. For instance, we do not know yet if the *s*-gram retrieval found relevant documents that were missed by the dictionary based approach. Likewise for data fusion, we do not know yet if we promoted into the final result set relevant results which were only present in some and not all of the experts used.

Acknowledgments

We are grateful to the AceMedia project (FP6-001765) which provided the image analysis tooklit. Research leading to this paper was supported by the European Commission under contract FP6-027026 (K-Space). The work of the first author is funded by Tampere Graduate School of Information Science and Engineering (TISE).

References

1. Grubinger, M., Clough, P., Hanbury, A., Müller, H.: Overview of the ImageCLEF-photo 2007 photographic retrieval task. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary (September 2007)

2. Hedlund, T., Keskustalo, H., Pirkola, A., Airio, E., Järvelin, K.: Utaclir @ CLEF 2001 - effects of compound splitting and n-gram techniques. In: CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, London, UK, Springer-Verlag (2002) 118–136
3. Hiemstra, D., Kraaij, W.: Twenty-One at TREC7: Ad-hoc and cross-language track. In: TREC. (1998) 174–185
4. Mcnamee, P., Mayfield, J.: Character n-gram tokenization for european language text retrieval. Information Retrieval **7**(1-2) (2004) 73–97
5. Järvelin, A., Järvelin, A., Järvelin, K.: s-grams: Defining generalized n-grams for information retrieval. Information Processing and Management **43**(4) (2007) 1005–1019
6. Pirkola, A., Keskustalo, H., Leppänen, E., Känsälä, A.P., Järvelin, K.: Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. Information Research **7**(2) (2002)
7. Keskustalo, H., Pirkola, A., Visala, K., Leppänen, E., Järvelin, K.: Non-adjacent digrams improve matching of cross-lingual spelling variants. In: SPIRE. (2003) 252–265
8. Markkula, M., Sormunen, E.: End-user searching challenges indexing practices inthe digital newspaper photo archive. Information Retrieval **1**(4) (2000) 259–285
9. AceMedia: The AceMedia Project, available at http://www.acemedia.org.
10. O'Connor, N., Cooke, E., le Borgne, H., Blighe, M., Adamek., T.: The AceToolbox: Low-Level Audiovisual Feature Extraction for Retrieval and Classification. In: 2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies. (2005)
11. Manjunath, B., Salembier, P., Sikora, T., eds.: Introduction to MPEG-7: Multimedia Content Description Language. Wiley (2002)
12. Belkin, N.J., Kantor, P., Fox, E.A., Shaw, J.A.: Combining the evidence of multiple query representations for information retrieval. Information Processing and Management **31**(3) (1995) 431–448
13. Wilkins, P., Ferguson, P., Smeaton., A.F.: Using score distributions for querytime fusion in multimedia retrieval. In: MIR 2006 - 8th ACM SIGMM International Workshop on Multimedia Information Retrieval. (2006)
14. Jarvelin, A., Wilkins, P., Adamek, T., Airio, E., Jones, G., Smeaton, A.F., Sormunen, E.: DCU and UTA at ImageCLEFPhoto 2007. In: ImageCLEF 2007 - The CLEF Cross Language Image Retrieval Track Workshop. (2007)
15. Robertson, S.: On score distributions and relevance. In: ECIR. (2007) 40–51
16. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language-model based search engine for complex queries (extended version). (2005-02-14 2005)
17. Pirkola, A.: The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In: SIGIR '98: Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval. (1998) 55–63
18. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In: CLEF. (2005) 652–661