Easing Erroneous Translations in Cross-Language Image Retrieval using Word Associations

Masashi Inoue

National Institute of Informatics, Tokyo, Japan
m-inoue@nii.ac.jp

Abstract. When short queries and short image annotations are used in text-based cross-language image retrieval, small changes in word usage due to translation errors may decrease the retrieval performance because of an increase in lexical mismatches. In the ImageCLEF2005 ad-hoc task, we investigated the use of learned word association models that represent how pairs of words are related to absorb such mismatches. We compared a precision-oriented simple word-matching retrieval model and a recall-oriented word association retrieval model. We also investigated combinations of these by introducing a new ranking function that generated comparable output values from both models. Experimental results on English and German topics were discouraging, as the use of word association models degraded the performance. On the other hand, word association models helped retrieval for Japanese topics whose translation quality was low.

1 Introduction

One of the goals of research on information retrieval (IR) systems is to overcome a shortage of meaningfully retrieved documents. In text-based ad-hoc image retrieval, when annotations are used as the target of query matching, an insufficient retrieval is often the result of term-mismatch. The words in a query do not appear in most annotations, because often there are few words in image annotations.

When a query and image annotations are described in different languages, and there needs to be translation process that brings diversity in the lexical expressions of a concept, a term-mismatch problem becomes more severe. As a result, the IR performance often degrades. In ImageCLEF2005, we studied the effect of word association models on mitigating such phenomena. We employed a probabilistic word-by-word query translation model structure [1], although in our models, the actual translation took place by an MT system outside of the retrieval model and the translation in the model was, in effect, a monolingual word expansion [2]. We tested our approach in the setting where both queries and annotations were short. Monolingual English-to-English, cross-lingual German-to-English, and cross-lingual Japanese-to-English image retrievals were compared.

One finding from our experiments was that when a simple word-matching strategy failed to retrieve a relevant image because of an erroneous translation, the use of a word association model could improve the word-matching. In our runs, a recovery effect was observed only in Japanese-to-English translations, being an example of translation between disparate languages.

In the following text, we first describe the experimental conditions, and then introduce the retrieval models and ranking functions. Next, we discuss the experimental results, and finally, we conclude the paper.

2 Data Preparation

2.1 Test Collection

The test collection used was the ImageCLEF2005 St Andrews Library photographic collection that was prepared for ad-hoc retrieval tasks [3]. This consisted of 28,133 images and their captions in English, with 28 topics in a variety of languages. Each caption had nine fields assigned by experts. Among these, we used only short title fields that were considered to be the simplest form of annotation. The mean length of the short titles was 3.43 words.

The retrieval topics was described using two fields: short description (title) and long description (narrative). They were the translations of original English topics. In our experiments, we used only the titles, which can be regarded as the approximation of users' queries. We examined English, German, and Japanese topics. The mean length of the queries was 4.18 words for English, 4.39 words for German, and 5.96 words for Japanese. We considered English topics as the baseline, German topics as the relatively easy task, and Japanese topics as the relatively hard task. Here, by 'easy' we mean that the current state-of-the art accuracy of machine translation (MT) for that language is high, and retrieval can be conducted in nearly the same fashion as the original (English) language. Similarly, by 'hard', we mean that queries differ substantially from the source language after undergoing the machine translation process. According to the results of ImageCLEF2004 that consisted of the same image dataset as ImageCLEF2005 but with different topics, German topics yielded the highest average mean average precision (MAP) score after English, and Japanese topics yielded the lowest average MAP scores for the top five systems [4].

The size of the vocabulary was 9,945 for both image annotations and queries. Although we were also interested in the use of visual information, we did not use it either for queries or for annotations. Therefore, the retrieval was purely textual. Details of data pre-processings are explained in [5].

2.2 Query Translation

Our approach to cross-language retrieval was to use query translation. According to previous experiments on ImageCLEF ad-hoc data, query translation generally outperforms document translation [6]. Although a combination of query translation and document translation may be promising, we only considered query

translation for now. German and Japanese topics were translated into English, the document language, using the Babelfish web-based MT system[1], and the complete list of translation results can be found in the Appendix of [5].

By analysing the translation results, we confirmed that German topics were 'easy' and Japanese topics were 'hard', in terms of the number of translation errors. In this paper, we define an error in machine translation as being the generation of words that invokes a mismatch between the queries and the annotations. For example, if a word is translated into 'photographs' when it should be translated to 'pictures', for a human observer, this difference has little effect in understanding sentences that contain the word 'photographs'. However, for image retrieval in particular, where only short text descriptions are available, such a difference may change the results of retrieval dramatically. For example, when all the relevant images are annotated as 'pictures', the system cannot retrieve anything, and therefore, this translation is considered an error. These errors can be observed only indirectly by comparing IR performances on the original topics and the translated topics. Therefore, in the following qualitative analysis, we only describe the errors that can be analysed qualitatively.

First, we examined the overall quality of German–English translations. Some notable errors were found in the translation of prepositions. For example, 'on' was translated as 'at', and 'from' was translated as 'of'. Other typical errors were the inappropriate assignment of imprecise synonyms. For example, 'ground' was replaced by 'soil'. (Details of the errors are given in [5].) Despite these errors, in most translations of German topics, the basic meanings were similar to the original English. Among 28 topics (titles), four topics were translated exactly as in the original English. This result confirms the relatively high accuracy of German–English MT.

For Japanese-to-English translations, however, the quality of translation was worse. As in the German-to-English translations, the Japanese-to-English translations contained errors in prepositions. Errors that were peculiar to the Japanese-to-English translations were the excessive use of definite articles and relative pronouns. More seriously, some of the Japanese words could not be translated at all. Untranslated words were 'aiona (Iona)', 'nabiku (waving)', and 'sentoandoryusu (St Andrews)'. The problem was that the untranslated words were often proper nouns, which can be useful for distinguishing relevant documents from irrelevant documents. Although this out-of-vocabulary problem occurred in German-to-English translations too, the effect of missing proper nouns was less severe, because the spellings were the same for both English and German, and for the indexing purposes, they did not need to be translated.

3 Retrieval Process after Translation

3.1 Retrieval Models

We introduce retrieval models based on the unigram language models and word association models. The baseline model was a simple unigram keyword-matching

[1] http://babelfish.altavista.com

document model denoted by `diag`. For the query of the length K, $\mathbf{q} = \{q_1, ..., q_K\}$, the likelihood of $\mathbf{q}$ being generated from $\mathbf{d}_n$, the nth document or image, is $\prod_{k=1}^{K} P(q_k|\mathbf{d}_n)$. Here, we assume independence between query words, $P(\mathbf{q}) = \prod_{k=1}^{K} P(q_k)$, although this is not always true for the ImageCLEF2005 topics, where titles are sometimes sentential and word orders have meaning. For the word association model, we estimated the following transitive probabilities from the jth word to the ith word in the vocabulary, $P(w_i|w_j)$. When the above two models are combined, the following represents the process of query generation:

$$\prod_{k=1}^{K} \sum_{i=1}^{V} P(q_k|w_i)P(w_i|\mathbf{d}_n). \tag{1}$$

The word association models can be estimated in various heuristic ways. We tried two methods, and in both methods, we regarded the frequency of the co-occurrence of two words as being the measure of word association. If two words co-occurred, then they were assumed to be related. The first method counted self-co-occurrences, where a word is regarded as co-occurring with itself as well as other co-occurrences. Values for each term pair were estimated as follows

$$P(w_i|w_j) = \frac{\#(w_i, w_j)}{\sum_{i=1}^{V} \#(w_i, w_j) + \#(w_i)} \qquad \text{where } i \neq j, \tag{2}$$

$$P(w_i|w_j) = \frac{\#(w_i, w_j) + \#(w_i)}{\sum_{i=1}^{V} \#(w_i, w_j) + \#(w_i)} \qquad \text{where } i = j. \tag{3}$$

Here, $\#(w_i, w_j)$ represents the frequency of co-occurrence of w_i and w_j (i.e., the appearance of the two words in the same image annotation), and $\#(w_i)$ represents the frequency of occurrence of w_i. This procedure strengthens self-similarities in the model and is termed `cooc`. The second method counted purely co-occurring pairs, and was named `coocp`. Values for each term pair were estimated as follows

$$P(w_i|w_j) = \frac{\#(w_i, w_j)}{\#(w_j)} \qquad \text{where } \#(w_j) > 0. \tag{4}$$

When we consider the matrix representations of above association probabilities, the baseline model that did not use a word association model can be interpreted as using an identity matrix and we denoted this as `diag`. Note that these models were estimated before the arrival of any queries and the computation at the time of query focused on score calculation.

3.2 Ranking Functions

Our runs were divided into two groups according to the ranking function employed. In the first group, documents were ranked according to the query–log likelihood of the document models. The ranking function can be written as

$$\log L = \sum_{k=1}^{K} \log \sum_{i=1}^{V} P(q_k|w_i)P(w_i|\mathbf{d}_n). \tag{5}$$

Runs based on these functions are marked with `log_lik` in Table 1.

In general, when an expansion method is involved, the number of terms matched between queries and documents increases. Consequently, the scores of documents given by the first scoring measure `log_lik` are larger in models with an expansion method than in those without an expansion method. Thus, the first scoring measure was not suitable for a comparison of the output scores between different models. The output combination method that will be introduced in Sect. 3.3 requires comparable scores from different models. Therefore, we heuristically derived the second measure. In the second group of runs, documents were ranked according to the accumulated information for all the matched words. First, we transformed the variables for the probability of a query word, q_k, $P(q)$, to $F_q = e^{(\log P(q))^{-1}}$ where $P(q)$ was either $P(q|\mathbf{d}_n)$ or $\sum_{i=1}^{V} P(q|w_i)P(w_i|\mathbf{d}_n)$, and was considered only when $P(q) \neq 0$. Then, the new ranking function can be defined as

$$\log L' = \sum_{k=1}^{K} \log \frac{1}{F_{q_k}}. \tag{6}$$

We regarded $\log \frac{1}{F_{q_k}}$ as the information on query word, q. A document with a higher score was assumed to have more information on the query than one with a lower score. Runs based on this measure are marked with `vt_info` in Table 1.

3.3 Model Output Combination

When the `vt_info` measure is used, the combination of different models at the output level can be performed because their scores are directly comparable. First, two sets of document scores and corresponding document indices from two models were merged. Then they were sorted in descending order of scores. For each document, the higher score was retained. This process assumed that lower scores usually corresponded to a lack of knowledge about the documents, and thus were less reliable. From the merged rank, the top M documents were extracted as the final result. This can be considered as an example of the raw score method [7]. Here, the scores are calculated by taking only matched terms into account. Strictly, this is not a single association model, however, for simplicity of notation, we denote it as `dc` association model to represent the combination of `diag` and `cooc`.

4 Experimental Results

The MAP scores in Table 1 are based on runs we conducted considering the 1,000 top scores for each of the 28 topics. On comparing our runs to those of other participants, the overall performance was found to be deficient. This is due to the restricted textual information we used and the oversimplification of our retrieval models and pre-processings. Because we were interested in a comparison

Table 1. Summary of the mean average precision scores (Figures in bold face represent the best performances for each language)

Ranking Function	log-lik			vt-info			
Association Model	diag	cooc	coocp	diag	cooc	coocp	dc
English	**0.0301**	0.0195	0.0065	0.0144	0.0110	0.0018	0.0149
German	**0.0215**	0.0077	0.0022	0.0110	0.0059	0.0064	0.0114
Japanese	0.0109	0.0120	0.0087	0.0118	0.0116	0.0078	**0.0121**

between query languages and the use of word association models, we will not discuss further the overall performance here.

First, we considered the difference between the models. In both English and German, our best run was achieved using the `diag` model, which we had considered as the simplest baseline. All models employing word association underperformed for these two languages. There are two possible explanations for this result. The first reason may be that there was no need to relax the limitation of exact term matching. Some relevant documents could be retrieved by word-by-word correspondence and other relevant documents could not be reached by word-level expansion. For example, the relevant images for topic 28 should be colour images. However, the textual description itself does not inform if an image is in colour or is monochrome. When such visual information is the dominant factor in determining relevance, changes in word-matching do not influence the retrieval results. The second reason may be that the word association models were not learned adequately, so they could not help with connecting query words and document words. Separation of the two types of influences in the final rankings is open to question.

For the model output combination method (`dc`), Figure 1 shows whether the `dc` or `diag` model performed better in monolingual English-English retrieval when the `vt_info` measure was used. The bars for each topic represent the difference between the average precision scores of two models on the top $1,000$ ranks. In Topics 11 and 25, the `dc` method worked better than the `diag` method did by taking advantages of the `cooc` method. Interestingly, Topics 11 and 25 that gave average precision gains in the `dc` model were not the most successful topics in `cooc`. For example, when the `cooc` model was used, Topic 2 benefited more. These results means that the gain achieved by the output combination was not simply derived by the quality of association model, but was provided by the merging process. Let us now look at the final ranking in detail. Figure 2 shows which of the two methods, `diag` or `cooc`, determined the position of the images in the merged ranking for monolingual (English-to-English) retrieval. The diamond symbols represent documents whose ranks were given by the precision-oriented `diag` models, and the square symbols represent documents whose ranks were given by the recall-oriented `cooc` models. Note that these figures do not hold information on the relevance of the documents, and the rank interlacing may have degraded the quality of the ranking. As can be seen in Figure 2, the

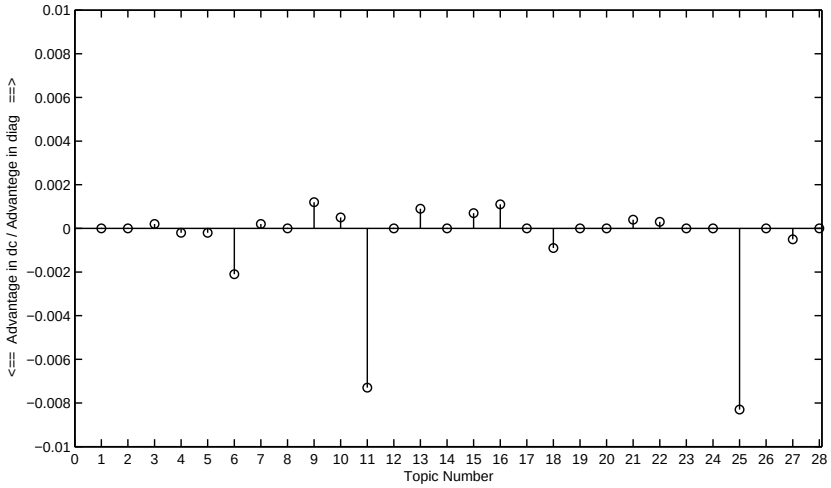

Fig. 1. Superiority of the two models in terms of resulting average precision for each topic (English-English retrieval evaluated on the top 1,000 ranks) when the `vt-info` measure was used

`diag` model dominated the top scores. We had expected this tendency, because an exact-matching scheme should have higher confidence in its outputs when queries can find their counterparts. What was unexpected was that in most of the topics, the dominance of the `diag` model often ranged from the top rank to about the 1,000th rank, and the scores given by `cooc` models appeared only in the lower ranks. Because we had considered only the top 1,000 ranks, the resulting MAP scores were determined almost solely by the `diag` model. Top-ranked documents are usually more important to a user, and with this in mind, we must consider a better way of rank merging so as not to miss any opportunity to swap top-ranked documents.

Next, we examined the effects of translations by comparing the three topic languages in baseline models. Basically, as we expected, monolingual topics performed best, German topics were in second place, and the performances of the Japanese topics were the worst. This order can be understood by the influence of translation errors, as discussed in Sect. 2.2. Particularly, the most serious problem in translation errors was the generation of out-of-vocabulary words. Most of the English topics after removal of any out-of-vocabulary words still made sense, whereas translated German and Japanese topics suffered from word scarcity. The table in Appendix A is the translation results of Japanese topics. It also shows which words were not contained in the target dataset or the short titles of images in our case. Note that, here by 'out-of-vocabulary', we mean unknown words for the IR models and not for the MT systems, as discussed in Sect. 2.2. The problem of these out-of-vocabulary words may be mitigated by using stemming, pseudo-relevance feedback, and use of external knowledge on

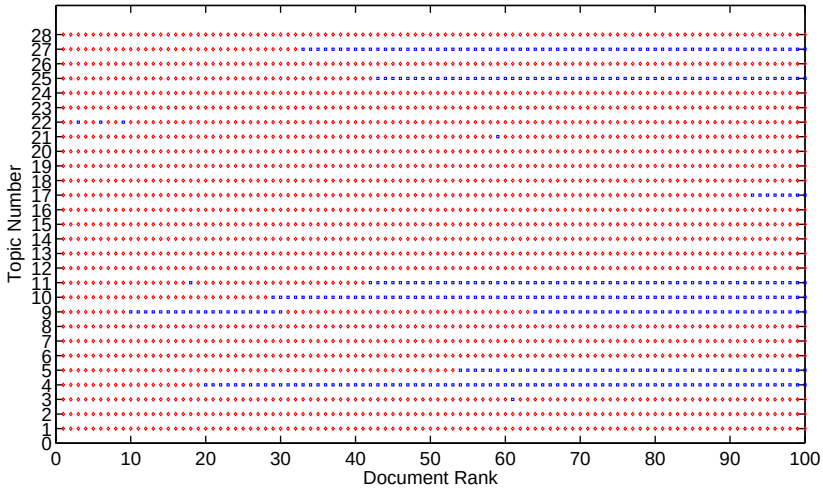

Fig. 2. Model dominance for the `dc` method in the top 100 scale (The diamond symbols represent documents whose ranks were given by the `diag` models, and the square symbols represent documents whose ranks were given by the `cooc` models)

the word associations. Investigation of their effect on the IR performance is a topic for future work.

Concerning the relationships between the topic languages and the association models, as we can see in Table 1, for the `log_lik` ranking function, direct word-matching models performed better than word association models in English and German topics. In contrast, in Japanese topics, the use of word association models (`cooc`) improved the performance. When English and Japanese topics were compared, because the only difference between languages was the presence or absence of translations, the positive effect of word association in Japanese topics may be attributed to the poor quality of translations. Therefore, word association models may be seen as the restoration of translation errors that caused mismatches in the retrieval process. When we also consider German topics, the relationship becomes more complex. Even though German topics contained some translation errors, the degradation of performance using `cooc` was more severe in German than in English. This result may be better understood by considering additional languages with various translation difficulties.

5 Discussion

In our experiments, we observed that the use of word association models may help recover query translation errors that arise in MT systems. However, the performances of our models were inadequate as standard systems. For simplicity, we did not incorporate the following established techniques: 1) inverse document frequency (idf) factor, 2) stop words elimination, and 3) document length

normalization. These may be integrated into the IR process to demonstrate the general applicability of our method.

There are other ways of utilizing word associations which may be of considerable benefit. We fixed the association models before the querying time. However, together with relevance feedback or pseudo-relevance feedback, association models can be estimated (e.g., [8]). Although the practicality of the construction of word association models from scratch is debatable, because the users' load may be too high, modification of already estimated associations at querying time using feedbacks will be an interesting extension of our approach. Another situation may arise when words are expanded more than once. In our runs, we used an MT system with a single output. If we had used an MT system that outputs multiple candidates with their confidence scores, then the MT system would have performed the soft expansion by itself. The combined effect of the expansion by the MT system and that by the IR system is an interesting future topic.

6 Conclusions

Text-based cross-language image retrieval that relies on short descriptions is considered to be less robust with respect to translation errors. In our experiments using the ImageCLEF2005 ad-hoc test collection, estimated word association models helped with the retrieval of Japanese topics when machine translation into English performed poorly. This recovery effect produced by word expansion may become clearer by comparing various languages with different degrees of translation difficulty.

References

1. Kraaij, W., de Jong, F.: Transitive CLIR Models. In: RIAO, Vaucluse, France (2004) 69–81
2. Inoue, M., Ueda, N.: Retrieving Lightly Annotated Images Using Image Similarities. In: SAC '05: Proceedings of the 2005 ACM symposium on Applied computing, NY, USA (2005) 1031–1037
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 Cross-language Image Retrieval Track. In: Proceedings of the Cross Language Evaluation Forum 2005. Springer Lecture Notes in Computer science (to appear)
4. Clough, P., Müller, H., Sanderson, M.: The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In: ImageCLEF2004 Working Note. (2004)
5. Inoue, M.: Recovering Translation Errors in Cross Language Image Retrieval by Word Association Models. In: ImageCLEF2005 Working Note. (2005)
6. Clough, P.: Caption vs. Query Translation for Cross-language Image Retrieval. In: ImageCLEF2004 Working Note. (2004)
7. Hiemstra, D., Kraaij, W., Pohlmann, R., Westerveld, T.: Translation Resources, Merging Strategies, and Relevance Feedback for Cross-language Information Retrieval. In: Lecture Notes in Computer Science. Volume 2069. (2001) 102–115
8. Xiang Sean Zhou, T.S.H.: Unifying Keywords and Visual Contents in Image Retrieval. IEEE Multimedia **9** (2002) 23–33

A Out-of-Vocabulary Words in Queries

This appendix contains the table of out-of-vocabulary and in-vocabulary words in the translated Japanese queries. Due to the limitation of space, we omit the tables for English and translated German queries. In the table below, the words in italic did not appear in the image annotations (out-of-vocabulary). Thus, only the words in bold face were effectively used. Note that our experimental procedure did not involve a stemming process and the presence of out-of-vocabulary words may be exaggerated.

Table 2. Translated Japanese queries

Topic No.	Translated Titles
1	*terrestrial airplane*
2	**the people** *who* **meet in the field music hall**
3	**the dog** *which sits* **down**
4	**the steam ship** *which* **is** *docked* **to the pier**
5	*image* **of** *animal*
6	*smallsized* **sailing ship**
7	**fishermen on boat**
8	**the building** *which* **the snow** *accumulated*
9	**the horse** *which* **pulls the load carriage and the carriage**
10	**photograph of sun Scotland**
11	**the Swiss mountain scenery**
12	**the** *illustrated postcards* **of Scotland and island**
13	**the** *elevated* **bridge of the** *stonework which* **is** *plural* **arch**
14	**people of market**
15	**the golfer** *who does* **the** *pad* **with the green**
16	**the** *wave which washes* **in the beach**
17	**the man or the woman** *who reads*
18	**woman of white dress**
19	*illustrated postcards* **of the** *synthesis* **of** *province*
20	**the Scottish visit of king family other** *than* **fife**
21	**poet Robert Burns' monument**
22	**flag building**
23	**grave inside church and large** *saintly* **hall**
24	*closeup* **photograph of bird**
25	**gate of arch** *type*
26	**portrait photograph of man and woman** *mixed* **group**
27	**the woman or the girl** *who has* **the** *basket*
28	*colour* **picture of forest scenery of every place**