

# Symmetric Statistical Translation Models for Automatic Image Annotation

Feng Kang and Rong Jin\*

## Abstract

Automatic image annotation provides means for users to search image collections on the semantic level using natural language queries. In the past, statistical machine translation models have been successfully applied to automatic image annotation. A problem with this approach is that, due to the skewed distribution of term frequency for annotation words, common words have been overly favored, which leaves little room for uncommon words to be used in auto-annotations. In contrast, studies on information retrieval have revealed that uncommon words are at least as important as common words since they are also often used in users' queries. Unlike the previous studies where a single type of statistical translation model is considered for automatic image annotation, in this paper, we studied two types of statistical translation models: a forward translation model, which *translates* visual information into textual words, and a backward model, which *translates* textual words into visual images. In particular, we propose a new statistical translation model, named regularization-based symmetric statistical translation model, which combines strength of forward and backward models to alleviate the problem of overly favoring common words. Our empirical studies with the Corel dataset have shown that the proposed model performs considerably better than the existing translation model and a state-of-the-art approach for automatic image annotation.

## 1 Introduction.

Efficient access to image databases requires the ability to search and organize images effectively. While images could be retrieved based on their features such as color, texture, it is usually more natural and desirable for users to search image databases using textual queries. One important reason is that textual queries allow users to express their information needs on the semantic level instead of the level of preliminary image features.

A key to image retrieval using textual queries is image annotation. Given annotated words for images, the problem of image retrieval becomes a problem of textual retrieval. Many well-developed textual retrieval algorithms, such as language modeling approaches [11,

12, 18, 19], can be applied to find images that are relevant to textual queries. Since manual annotation is usually expensive and subjective, many methods have been developed to annotate images automatically [1-3, 5, 6, 8, 10, 13-17].

A statistical machine translation model for automatic image annotation [8] views the process of annotating images as a process of translating information from a 'visual language' to textual words. Images are first segmented into different regions, which are further grouped into a number of clusters, or image blobs as in [8]. Then, correspondence between image blobs and annotated words is learned through a statistical machine translation model [4].

One difficulty with translation models for automatic image annotation arises from the skewed distribution of word frequency. According to [4], a key for translation models to disambiguate the alignment between image regions and annotated words is the co-occurrence statistics. If an image blob co-occurs more frequently with a word 'A' than with any other words, it will be more likely for the image blob to be associated with 'A'. According to [16], the term frequency of annotation words follows the Zipf's law, namely a small number of words appear very often and most words are used only by a few images. As a result, a common word can *accidentally* co-occur with a blob that in fact is more associated with an uncommon word. The problem with the co-occurrence statistics is further complicated by the fact that massive number of image regions are first clustered into a small number of blobs. Very often, image regions for different annotation words have similar distributions over the space of image features and thus are clustered into the same group (i.e., image blob). As a result, an image region related to a rare word could be grouped with other image regions related with common words, which leads to more errors in co-occurrence statistics.

To correct the potential errors in the co-occurrence statistics, we examine two types of translation models. Most previous studies on translation models for automatic image annotation focus on the model that *translates* image regions/blobs into textual words, which is called *forward translation model* in this paper. Apparently, we can apply the translation model in a reverse way, namely translating textual words into im-

\*Department of Computer Science and Engineering, Michigan State University, East Lansing MI 48824. {kangfeng, rongjin}@cse.msu.edu

	w1	w2
b1	100	50
b2	200	35
Forward translation model $p(w b)$		
	w1	w2
b1	0.67	0.33
b2	0.85	0.15
Backward translation model $p(b w)$		
	w1	w2
b1	0.33	0.58
b2	0.67	0.42

Table 1: An example of forward and backward translation models for two image blobs(b1 and b2), and two words (w1 and w2).

age blobs, which we call a *backward translation model*. These two kinds of translation models make different assumptions between blobs and words. The forward translation model assumes that each image blob is translated into a single word, while each word can be translated into multiple blobs. The backward translation model is based on the assumption that each word is translated into a single image blob.

In order to better illustrate the difference between these two types of translation models, consider a simple example of co-occurrence statistics shown in Table1. On one hand, for the forward translation model, the translation probabilities for word ‘w1’ are dominative for both image blobs. It is unlikely for word ‘w2’ to be used in any auto-annotations. On the other hand, for the backward translation model, we do find that image blob ‘b1’ is strongly associated with word ‘w2’ and the chance for image blob ‘b2’ to be associated with word ‘w2’ is also high. The difference motivates us to propose a symmetric model, which combines the two models together.

The rest of this paper is arranged as follows: Section 2 describes the background knowledge; Section 3 describes the proposed symmetric translation model that combines the forward and backward translation models; The empirical studies are described in Section 4; Section 5 draws the conclusions.

## 2 Related Work.

In this section, we describe the overview of statistical methods for automatic image annotation with the focus on translation model approaches.

**2.1 Automatic Image Annotation.** The key to automatic image annotation is to learn the annotation

models that automatically predict annotation words given extracted image features. A variety of machine learning methods have been applied to automatic image annotation, including machine translation model [8], co-occurrence model [17], latent space approaches [1, 16], graphic models [3], classification approaches [5, 6, 14], and relevance language models [10, 13]. The co-occurrence model [17] collects the co-occurrence counts between words and image features and uses them to predict annotated words for images. Duygulu et al. [8] improved the co-occurrence model by utilizing machine translation models. Another way of capturing co-occurrence information is to introduce latent variables that link image features with words. Methods in this category include latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA) [16], hierarchical aspect model [1], Gaussian Mixture Model (GMM), Latent Dirichlet Allocator (LDA), and correspondence LDA [3]. The classification approaches for automatic image annotation treat each annotated word as an independent class and create a different image classification model for every word. Work such as linguistic indexing of pictures [14], image annotation using SVM [6] and Bayes point machine [5] fall into this category. More recently, relevance language models have been applied to automatic image annotation [10, 13]. Empirical studies [10, 13] have shown that relevance language models for image annotation are better than translation models.

**2.2 Machine Translation Models for Automatic Image Annotation.** Using the IBM translation model I [4, 8], the probability of annotating image blobs  $\vec{b}_i = \{b_{i,1}, b_{i,2}, \dots, b_{i,m}\}$  with words  $\vec{w}_i = \{w_{i,1}, w_{i,2}, \dots, w_{i,n}\}$ , i.e.,  $p(\vec{w}_i | \vec{b}_i)$ , is expressed as follows:

$$\begin{aligned}
 p(\vec{w}_i | \vec{b}_i) &= \prod_{j=1}^n p(w_{i,j} | \vec{b}_i) \\
 (2.1) \quad &\propto \prod_{\{j|w_{i,j}=1\}} \sum_{k=1}^m t_{j,k} b_{i,k}
 \end{aligned}$$

where  $t_{i,j}$  stands for the probability of translating the  $k$ -th blob into the  $j$ -th word. In order to annotate an image  $I = \{b_1, b_2, \dots, b_m\}$ , (2.1) is applied to find the set of words  $\vec{w}$  that maximizes  $p(\vec{w} | \vec{b})$ . The translation probabilities  $\{t_{j,k}\}$  can be obtained by maximizing the log-likelihood of training images, i.e.,

$$\begin{aligned}
 \log l(T) &= \sum_{i=1}^{|T|} p(\vec{w}_i | \vec{b}_i) \\
 (2.2) \quad &= \sum_{i=1}^{|T|} \sum_{j=1}^n w_{i,j} \log \left( \sum_{k=1}^m t_{j,k} b_{i,k} \right)
 \end{aligned}$$

Expectation-Maximization (EM) algorithm [7] is applied to find the optimal solution for (2.2) and updating equation is:

$$(2.3) \quad t_{j,k}^{new} = \frac{1}{Z_k} \sum_i \frac{w_{i,j} b_{j,k} t_{j,k}^{old}}{\sum_{k'} w_{j,k'} t_{j,k'}^{old}}$$

$Z_k$  is a normalization factor that ensures  $\sum_j t_{j,k}^{new} = 1$ , namely each blob has to be translated into a single annotation word. Detailed EM algorithm is described in [4]. The above translation model takes the direction of translating image blobs into words, which we call *forward translation model*. We can take another direction of translation, i.e., translating words into blobs, which we call *backward translation model*. In backward translation model, the translation probability from annotation words to image blobs is written as:

$$(2.4) \quad p(\vec{b}_i | \vec{w}_i) = \prod_{j=1}^m p(b_{i,j} | \vec{w}_i) \propto \prod_{j=1}^m \left\{ \sum_{k=1}^m u_{j,k} w_{i,k} \right\}^{b_{i,j}}$$

Similarly, EM is used to find the set of translation probabilities and the updating equation is written as:

$$(2.5) \quad u_{j,k}^{new} = \frac{1}{Z_k} \sum_i \frac{w_{i,j} b_{i,k} u_{j,k}^{old}}{\sum_{k'} w_{i,k'} u_{j,k'}^{old}}$$

where  $u_{j,k}$  stands for the probability of translating the  $k$ -th word into the  $j$ -th blob.  $Z_k$  is a normalization factor that ensures  $\sum_j u_{j,k}^{new} = 1$ , namely each word has to be translated into a single image blob.

In next section, we propose a new translation model, which combines the forward and backward translation models to enhance the quality of automatic image annotations.

### 3 Regularization-based Symmetric Translation Model: Combining the Forward and Backward Translation Models.

The main idea of the proposed model, a **regularization-based symmetric translation model (RSTM)**, is first to examine the discrepancy between the forward and backward models, and then to correct them by utilizing the information across the two models. We first introduce a symmetric KL divergence term that measures the discrepancy between the forward and backward models:

$$(3.6) \quad KL = \sum_j \sum_k p(w_j, b_k; f) \log\left(\frac{p(w_j, b_k; f)}{p(w_j, b_k; b)}\right) + \sum_j \sum_k p(w_j, b_k; b) \log\left(\frac{p(w_j, b_k; b)}{p(w_j, b_k; f)}\right)$$

According to the property of KL divergence, the above expression becomes zero iff  $p(w_j, b_k; f) = p(w_j, b_k; b)$  for any  $j \in [1..n]$  and  $k \in [1..m]$ . Then, we add the KL divergence term into the objective function as the regularization term to ensure the consistency between the forward and backward translation models:

$$(3.7) \quad \Omega_{RSTM} = \left\{ \sum_{i=1}^{|T|} \sum_{\{w_{i,j}=1\}} \log\left(\sum_{k=1}^m t_{j,k} b_{i,k}\right) + \underbrace{\sum_{i=1}^{|T|} \sum_{\{b_{i,k}=1\}} \log\left(\sum_{j=1}^n u_{k,j} w_{i,j}\right)}_{\text{translation}} - \lambda \left\{ \sum_j \sum_k p(w_j, b_k; f) \log\left(\frac{p(w_j, b_k; f)}{p(w_j, b_k; b)}\right) + \sum_j \sum_k p(w_j, b_k; b) \log\left(\frac{p(w_j, b_k; b)}{p(w_j, b_k; f)}\right) \right\} \right\}_{\text{regularization}}$$

where  $\lambda$  is that determine the degree of consistency between the two models. Efficiently finding the optimal solution to (3.7) is more complicated than the EM algorithm for standard translation model. Here, we list the updating equations for the forward and backward translation models, and leave out the detailed derivation for brevity.

$$(3.8) \quad t_{j,k}^{new} = \frac{2C_{j,k}}{B_{j,k} + \sqrt{B_{j,k}^2 + 4A_{j,k}C_{j,k}}}$$

$$A_{j,k} = 2\lambda \frac{p(b_k)}{t_{j,k}^{old}}$$

$$B_{j,k} = \lambda p(b_k) \log(t_{j,k}^{old}) - \lambda p(b_k) \log(u_{k,j} p(w_j)) + \alpha_k$$

$$C_{j,k} = \sum_{i=1}^{|T|} \frac{t_{j,k}^{old} b_{i,k}}{\sum_{k'=1}^m t_{j,k'}^{old} b_{i,k'}} + \lambda p(w_j) u_{k,j}$$

where  $\alpha_k$  is the normalization factor that ensures  $\sum_j t_{j,k}^{new} = 1$ .

$$(3.9) \quad u_{k,j}^{new} = \frac{2F_{k,j}}{E_{k,j} + \sqrt{E_{k,j}^2 + 4D_{k,j}F_{k,j}}}$$

$$D_{k,j} = 2\lambda \frac{p(w_j)}{u_{k,j}^{old}}$$

$$E_{k,j} = \lambda p(w_j) \log(u_{k,j}^{old}) - \lambda p(w_j) \log(t_{j,k} p(b_k)) + \beta_j$$

$$F_{k,j} = \sum_{i=1}^{|T|} \frac{u_{k,j}^{old} w_{i,j}}{\sum_{j'=1}^n u_{k,j'}^{old} w_{i,j'}} + \lambda p(b_k) t_{j,k}$$

where  $\beta_j$  is the normalization factor that ensures  $\sum_k u_{k,j}^{new} = 1$ .

Probabilities  $p(w)$  and  $p(b)$  are used for joint probability  $p(w, b)$ . One natural choice for  $p(w)$  and  $p(b)$  is to use the empirical values that are estimated from the training corpus. However, one problem is that, the empirical distribution for term frequency follows a skewed distribution (Zipf’s law). As a result, if the empirical  $p(w)$  is used directly, the regularization term will mainly focus on the consistency checking for common words. For rare words, since its empirical  $p(w)$  is very small, its impact in the regularization term is almost ignorable. To put equal emphasis on both common words and uncommon words, we decide to use a uniform distribution for both  $p(w)$  and  $p(b)$ , which turns out to have better performance in our empirical studies.

## 4 Experiments.

In the following experiments, we compare the effectiveness of the proposed model to the existing translation model and a state-of-art statistical model for automatic image annotation.

**4.1 Experiment Data.** The same subset of Corel data used in [8] is used in this experiment. It consists of 5000 annotated images, among which 4500 of them are used for training and selection of parameters and the rest 500 images used for testing. 371 different words are used for annotating both training and testing images. Similar to the previous studies on automatic image annotation, the quality of automatic image annotation is measured by the performance of retrieving auto-annotated images regarding to single-word queries. For each single-word query, **precision** and **recall** are computed using the retrieved lists that are based on the true annotations and the auto-annotations. Let  $I_j$  be a test image,  $t_j$  be its true annotation, and  $g_j$  be its auto-annotation. For a given query word  $w$ , precision and recall are defined respectively as:

$$precision(w) = \frac{|\{I_j | w \in t_j \wedge w \in g_j\}|}{|\{I_j | w \in g_j\}|}$$

$$recall(w) = \frac{|\{I_j | w \in t_j \wedge w \in g_j\}|}{|\{I_j | w \in t_j\}|}$$

The  $precision(w)$  measures the accuracy in annotating images with word  $w$  and the  $recall(w)$  measures the completeness in annotating images with word  $w$ . The average precision and recall over different single-word queries are used to measure the overall quality of automatically generated annotations. The third metric, **#Ret\_Query**, is the number of single-word queries for

	TM	RM	RSTM
#Ret_Query	63	76	86
Average recall	0.2106	0.2656	0.3373
Average precision	0.1836	0.2299	0.2141

Table 2: Performance comparison of different models: the Translation Model(TM), the Relevance Model(RM), and the Regularization-based Symmetric Translation Model(RSTM).

which at least one relevant image can be retrieved:

$$\#Ret\_Query = |\{w | precision(w) > 0 \wedge recall(w) > 0\}|$$

This metric compensates the metrics of average precision and average recall by providing information about how wide is the range of words that contribute to the average precision and recall.

### 4.2 RSTM VS. Original Translation Model.

In this section, we compare the performance of the proposed translation model to the original one. First, we employ the cross-validation approach to decide the best value for  $\lambda$ . In particular, the training data is divided into two parts: 70% of data is used for training the model, and 30% of data is used for cross validating the value of  $\lambda$ . We find  $\lambda = 50,000$  is a good choice for the RSTM.

The comparison result is listed in Table 2. The RSTM performs substantially better than the original translation model in all three metrics. The most difference between them is in the metric of #Ret\_Query and average recall, which is 86 and 33.73% for the RSTM, and is only 63 and 21.06% for the original translation model. Some examples of the annotation words are listed in Table 3. We can see that, the RSTM is able to come up with more specific words for annotation than the original translation model. For instance, consider the first example in Table 3, the original translation model is only able to come up with a general/common term ‘buildings’ to describe the object in the image, while the RSTM is able to identify the building as ‘lighthouse’.

### 4.3 Comparison to Other Annotation Models.

We also compare RSTM to the relevance language model for automatic image annotation, which has shown good performance in recent studies [10, 13]. The result is listed in Table 2. Compared to the RSTM, the relevance language model achieves slightly better performance in terms of average precision. However, its #Ret\_Query and average recall are lower than the RSTM, with 76 and 26.56% versus 86 and 33.73% for the RSTM.

Image	TM	RSTM	Manual
	sky water tree buildings rocks	sky water rocks booby lighthouse	water hills coast lighthouse
	sky water tree buildings snow	water tree snow forest zebra	tree snow forest coyote
	sky water tree grass rocks	sky rocks fox giraffe tusks	rocks fox kit baby
	water tree people grass buildings	tree field horses albatross foals	field horses mare foals

Table 3: Examples of annotations generated by the Translation Model (TM), the Regularization-based Symmetric Translation Model (RSTM). The manual annotations are included in the last column.

## 5 Conclusion.

In this paper, we propose a regularization-based symmetric translation model (RSTM) to explore the correlation between the forward and the backward translation models. The RSTM introduces a soft regularization term based on the measurement of KL divergence. Empirical studies have shown that the model is able to effectively enhance the quality of automatic image annotations.

## References

- [1] K. Barnard, P. Duygulu and D. Forsyth, *Clustering Art*. in Proceedings of the IEEE Computer Society Conference on Pattern Recognition. 2001.
- [2] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei and M. I. Jordan, *Matching Words and Pictures*. Journal of Machine Learning Research, 2003. 3: p. 1107-1135.
- [3] D. Blei and M. Jordan, *Modeling annotated data*. in Proceedings of 26th International Conference on Research and Development in Information Retrieval (SIGIR). 2003.
- [4] P. Brown, S. D. Pietra, V. D. Pietra and R. Mercer, *The Mathematics of Statistical Machine Translation*. Computational Linguistics, 1993. 19(2): p. 263-311.
- [5] E. Chang, K. Goh, G. Sychay and G. Wu, *CBSA: content-based soft annotation for multimodal image retrieval using bayes point machines*. CirSysVideo, 2003. 13(1): p. 26-38.
- [6] C. Cusano, G. Ciocca and R. Schettini, *Image annotation using SVM*. in Proceedings of Internet imaging IV, Vol. SPIE 5304. 2004.
- [7] A. P. Dempster, N. M. Laird and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*. Journal of Royal Statistical Society, 1977. 39(1): p. 1-38.
- [8] P. Duygulu, K. Barnard, N. d. Freitas and D. A. Forsyth, *Object recognition as machine translation: learning a lexicon for a fixed image vocabulary*. in Proceedings of 7th European Conference on Computer Vision. 2002.
- [9] S. F. Chen, and J. Goodman, *An empirical study of smoothing techniques for language modeling*. in Annual Meeting of the ACL Proceedings of the 34th conference on Association for Computational Linguistics. 1996. Santa Cruz, California.
- [10] J. Jeon, V. Lavrenko and R. Manmatha, *Automatic Image Annotation and Retrieval using Cross-Media Relevance Models*. in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. 2003.
- [11] R. Jin, C. X. Zhai and A. G. Hauptmann, *Title Language Model for Information Retrieval*. in Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2002.
- [12] V. Lavrenko and B. Croft, *Relevance-based language models*. in The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001.
- [13] V. Lavrenko, R. Manmatha and J. Jeon, *A Model for Learning the Semantics of Pictures*. in Proceedings of Advance in Neural Information Processing. 2003.
- [14] J. Li and J. Z. Wang, *Automatic linguistic indexing of pictures by a statistical modeling approach*. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2003. 25(19): p. 1075-1088.
- [15] O. Maron, *Learning from Ambiguity*. 1998, MIT.
- [16] F. Monay and D. Gatica-Perez, *On Image Auto-Annotation with Latent Space Models*. in Proc. ACM International Conference on Multimedia. 2003.
- [17] Y. Mori, H. TAKAHASHI and R. Oka, *Image-to-Word Transformation Based on Dividing and Vector Quantizing Images With Words*. in MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management. 1999.
- [18] J. Ponte, *A Language Modeling Approach to Information Retrieval*. in Department of Computer Science. 1998, Univ. of Massachusetts at Amherst.
- [19] Zhai, C. X. and J. Lafferty. *Model-based feedback in the KL-divergence retrieval model*. in Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM). 2001.