Systematic Evaluation of Machine Translation Methods for Image and Video Annotation

Paola Virga[1] and Pınar Duygulu[2]

[1] Department of Computer Science, Johns Hopkins University, Baltimore, USA
[2]Department of Computer Engineering, Bilkent University, Ankara, Turkey
paola@jhu.edu,duygulu@cs.bilkent.edu.tr

Abstract. In this study, we present a systematic evaluation of machine translation methods applied to the image annotation problem. We used the well-studied Corel data set and the broadcast news videos used by TRECVID 2003 as our dataset. We experimented with different models of machine translation with different parameters. The results showed that the simplest model produces the best performance. Based on this experience, we also proposed a new method, based on cross-lingual information retrieval techniques, and obtained a better retrieval performance.

1 Introduction

With the recent developments in technology, there is a huge amount of digital multimedia data available in many archives and on the Internet. In order to efficiently and effectively access and make use of this huge amount of information, the automatic retrieval and annotation of multimedia data should be provided. This can be only achieved with the association of low-level and mid-level features with higher-level semantic concepts. However, this is a very difficult and long-standing problem and requires carefully labeled data, which is very difficult to obtain in large quantities.

Recently, it is shown that, such relationships can be learned from multimodal datasets that provide a loosely labeled data in large quantities. Such data sets include photographs annotated with a few keywords, news photographs on the web and videos with speech transcripts. With careful use of such available data sets, it is shown that semantic labeling of images is possible [1–3]. More recently, probabilistic models are proposed to capture the joint statistics between image regions and caption terms. These include the simple co-occurrence model [4], hierarchical aspect model [5], cross-media relevance model (CMRM) [6], Correlation Latent Dirichlet Allocation (LDA) model [7], and translation model [8].

In [8], Duygulu et.al. considers the problem of learning the correspondences between image regions and words as a translation process, similar to the translation of text in two different languages. The correspondences between the image regions and the concepts are learned, using a method adapted from Statistical Machine Translation. Then, these correspondences are used to predict words corresponding to particular image regions or to automatically annotate the images.

In this study, we analyze the machine translation approach for image annotation. Although, better results are reported in the literature, this method is simple and can be easily adapted to other applications. Also, it is shown that, when integrated to an information retrieval task, it produces the best results compared to some other methods [**?**]. Our goal is to provide a systematic evaluation of the machine translation approach and investigate the effect of different extensions to the basic model.

In [8], statistical machine translation idea is used in its simplest form. We experimented several other models and parameters of statistical machine translation methods and compare the results with the results of the simplest model. We also integrated the language modeling in the form of word co-occurrences. The results are evaluated on Corel and TRECVID 2003 data sets.

Also, as new method cross-lingual information retrieval CLIR techniques are adapted and shown that the retrieval performance is increased by the new proposed method.

The paper is organized as follows. First, the motivation for the machine translation approach will be given in Section 2. We will describe the data set in Section 3. The details of the basic approach will be presented in 4. Then, in Section5 we will present the experiments performed to analyze the machine translation approach. The results of applying CLIR techniques will be discussed in Section 6.

2 Motivation

In the image and video collections, the images are usually annotated with a few keywords which describe the images. However, the correspondences between image regions and words are unknown(Figure 1-a). This correspondence problem is very similar to the correspondence problem faced in statistical machine translation literature (Figure 1-b).

Brown *et.al* [10] suggested that it may be possible to construct automatic machine translation systems by learning from large datasets (aligned bitext) which consist of many small blocks of text in both languages, corresponding to each other at paragraph or sentence level, but not at the word level. Using these aligned bitexts, the problem of lexicon learning is transformed into the problem of finding the correspondences between words of different languages, which can then be tackled by machine learning methods.

Due to the similarity of problems, correspondence problem between image regions and concepts can be attacked as a problem of translating visual features into words, as first proposed by Duygulu *et.al.* [8]. Given a set of training images, the problem is to create a probability table that associates words and visual features which can be then used to find the corresponding words for the given test images.

3 Data Sets

In this study, we use Corel stock photos since that is a highly experimented data set for image annotation. We also incorporate the TREC Video Retrieval Evaluation (TRECVID) 2003 data set which consists of more than 100 hours of ABC and CNN broadcast news

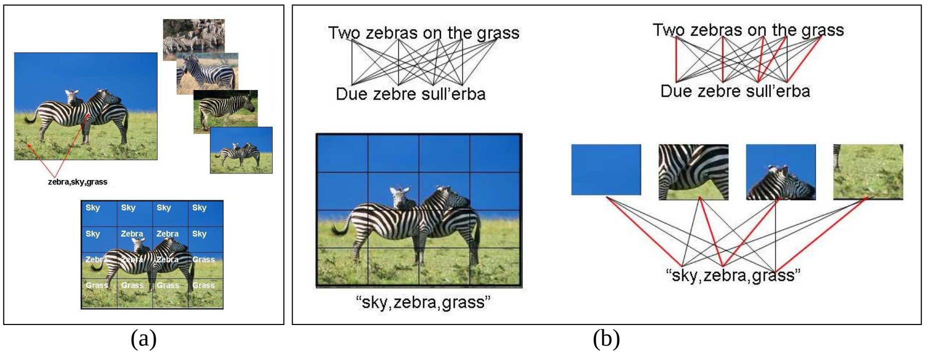


Fig. 1. (a)The correspondence problem between image regions and words. The words `zebra`, `grass` and `sky` are associated with the image, but the word-to-region correspondences are unknown. If there are other images, the correct correspondences can be learned and used to automatically label each region in the image with correct words or to auto-annotate a given image. (b) The analogy with the statistical machine translation. We want to transform one form of data (image regions or English words) to another form of data (concepts or French words).

videos [11]. For TRECVID dataset, the keyframes extracted from video are used as the images and the concepts manually annotated by the participants are used as the keywords to make the analogy to Corel data. For the experiments on Corel data, we use 4500 images for training and 500 images for testing. The number of annotation keywords is 374. For TRECVID dataset around 44K images are annotated by 137 concepts. We use 38K of the data for training and use a reduced set of 75 concepts with higher frequencies. The regions could be obtained by a segmentation algorithm as in [8], but in this study we prefer to use fixed sized blocks due to the simplicity and because of the more successful results reported in the literature. Corel images are divided into 24 rectangular blocks as used in [6], and from each block color and texture features are extracted. TRECVID keyframes are divided into 35 blocks, which are then represented by color, texture and edge features. For the TRECVID data we also experimented extracting features around interest points obtained a Harris corner detector based algorithm.

4 Basic Approach

In machine translation, a lexicon links a set of discrete objects (words in one language) onto another set of discrete objects (words in the other language). Therefore, in order to exploit the analogy with machine translation, both the images and the annotations need to be broken up into discrete items. The annotation keywords, which will be called as **concepts** can be directly taken as discrete items. However, visual data is represented as a set of feature vectors. In order to obtain the discrete items for visual data, the features are classified by vector quantization techniques such as K-means. The labels of the classes are then used as the discrete items for the visual data and called as **visterms**.

For TRECVID data the feature vectors are separately quantized into 1000 visterms each. For Corel data 500 visterms are obtained by using all the features at once.

The aligned bitext, consisting of the visterms and the concepts are used to construct a probability table linking visterms with concepts. Probability tables are learned using Giza$++$ [16], which is a part of Statistical Machine Translation toolkit developed during summer 1999 at CLSP at Johns Hopkins University.

Brown *et. al.* [10] propose a set of models for statistical machine translation (SMT). The simplest model (Model 1), assumes that all connections for each French position are equally likely. In the work of Duygulu et. al. [8], this model is adapted to translate visterms to concepts, since there is no order relation among the visterms or concepts in the data. As the basic approach, we also use Model 1 in the form of direct translation.

In order to annotate the images, the word posterior probabilities supplied by the probability table are used. The word posterior probabilities for the whole image are obtained by marginalizing the word posterior probabilities of all the visterms in the image:

$$P_0(c|d_v) = 1/|d_v| \sum_{v \in d_v} P(c|v) \tag{1}$$

where v is a visterm, d_v is the set of all visterms of the image and c is a concept. Then, the word posterior probabilities are normalized. The concepts with the highest posterior probabilities are used as the annotation words.

Figure 2-a shows some auto-annotation examples for Corel data. Most of the words are predicted correctly and most of the incorrect matches are due to the missing manual annotations (*e.g.* although `tree` is in the image on the top-left example it is not in the manual annotations). In Figure 2-b, the annotation results are presented for some images from TRECVID data by showing the concept which is predicted by the highest probability and matches with the manual annotations.

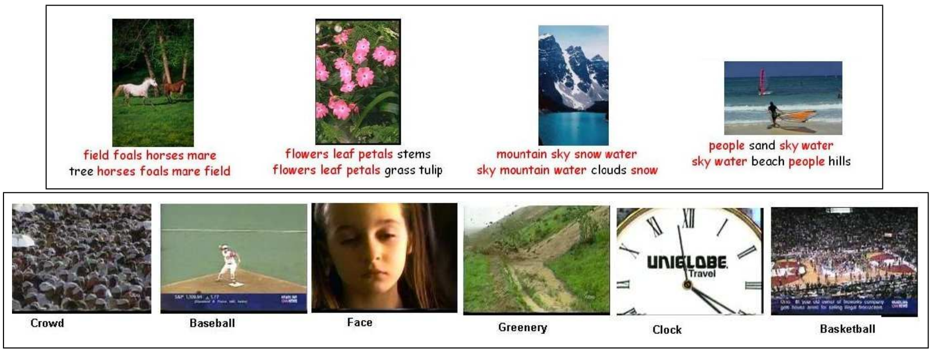

field foals horses mare
tree horses foals mare field
flowers leaf petals stems
flowers leaf petals grass tulip
mountain sky snow water
sky mountain water clouds snow
people sand sky water
sky water beach people hills
Crowd
Baseball
Face
Greenery
Clock
Basketball

Fig. 2. Annotation examples **top:** on Corel data set, and **bottom** on TRECVID data set. For Corel set the manual annotations are shown at the top, and the predicted words (top 5 words with the highest probability) are shown at the bottom. Words in red color correspond to the correct matches. For TRECVID data set, the concepts predicted with the highest probability and match with one of the annotation concepts are shown.

These annotation examples are obtained by using the features extracted from the rectangular blocks. For TRECVID data set, we also experimented the features extracted around the interest points. Figure 3 shows the effect of different features and compares the features extracted from the blocks and around interest points. It is observed that, the performance is always better when features are extracted from blocks. The experiments also show that, color feature gives the best performance when used individually but using a combination of all three features gives the best performance. The face information is also integrated in the form of the number of detected faces. However, this extra information did not give any significant improvement. Feature selection based on Information Gain is also experimented, but the results were not satisfactory. Based on these observations, in the rest of the experiments we prefer to use the combination of color, texture and edge features extracted from blocks.

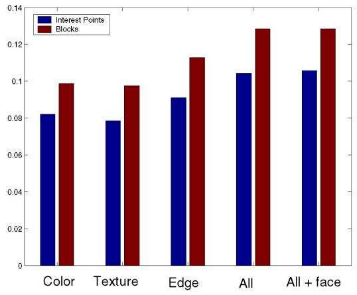


Fig. 3. Comparison between block-based features (red) and Harris interest point features (blue).

5 Analysis of the Machine Translation Approach

In this section we will analyze the machine translation approach by providing the results of different extensions to the basic approach. The results will be compared using the Mean Average Precision (mAP) values.

First, we experimented the effects of using higher models. We have trained our system with more complicated models: (i) using Model 2,(ii) HMM Model on top of Model 1 and, (iii) Model 4 on top of Model 1 and HMM Model training. However, the experiments show that, the simplest model (Model 1) results in the best annotation performance. The Mean Average Precision values obtained by Model 1 are 0.125 on the Corel data set and 0.124 on the TRECVID data set.

It is also observed that, the number of iterations in Giza++ training affects the annotation performance. Although, annotation performance decreases with the increased number of iterations, with less iterations less number of words can be predicted. Due to this tradeoff, number of iterations is set to 5 in the experiments.

We also incorporate the language modeling in the form of word cooccurrences, since our data sets consist of individual concepts without any order. In our new model, the probability of a concept given an image depends both to the probability of that concept given other concepts, and the probability of other concepts given the image.

$$P_1(c_i|d_v) = \sum_{j=1}^{|C|} P(c_i|c_j) P_0(c_j|d_v) \tag{2}$$

It is shown that (Table 1) incorporating word cooccurrences into the model helps to improve annotation performance for Corel data set, but does not create a difference for TRECVID data set.

Table 1. The effect of incorporating word co-occurrences.

	Corel	TRECVID
Model 1	0.125	0.124
Model 1 with word cooccurrences	0.145	0.124

Another experiment that has been studied but not performing well was using the alignments provided by training to construct a co-occurrence table. For this experiments we have trained Giza++ in both ways, i.e. one table is created for co-occurrences by training from visterms to concepts and another one is created by training from concepts to visterms. A third co-occurrence table is created by summing up the two tables. As shown in Table 2, the results were worse than the base results.

Table 2. Comparison of the results obtained from a co-occurrence table of the alignment counts with the basic Model 1 results. V represents visterms and C represents concepts.

Model 1	Alignment(V to C)	Alignment(C to V)	Alignment(Combined)
0.125	0.103	0.107	0.114

We will now review the IBM and the HMM translation models and their underlying assumptions, and argue why a more powerful translation model does not necessarily result in a better performance under MAP.

$$P(f|e) = \sum_{\mathbf{a}} P(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \sum_{\mathbf{a}} P(m|\mathbf{e}) P(\mathbf{a}|m, \mathbf{e}) P(\mathbf{f}|\mathbf{a}, m, \mathbf{e}) \tag{3}$$

where $P(f|e)$ is probability of translating the English sentence “**e**” of length l into the French sentence “**f**” of length m, and “**a**” represents the alignment between the two sentences. The following assumptions are made Model 1:

- $P(m|\mathbf{e}) = \epsilon(m|l)$ string length probabilities

- $P(\mathbf{a}|m,\mathbf{e}) = (l+1)^{-m}$ alignment probabilities
- $P(\mathbf{f}|\mathbf{a},m,\mathbf{e}) = \prod_{j=1}^{m} t(f_j|e_{a_j})$ word translation probabilities.

Model 2 differs from Model 1 in having the alignment probability in which the alignment a_j depends on j, l, m; more specifically $P(\mathbf{a}|m,\mathbf{e}) = \prod_{j=1}^{m} p(a_j|j,l,m)$. However, when working with concepts and visterms, we observe that the concept in the caption are not written in any particular order. For example, the blocks associated with sun and sky are always adjacent but the corresponding concept sentences can be annotated with any of the following word orders: $\{$*sky,sun,*$\cdots\}$, $\{$*sun,*$\cdots$*,sky* $\}$, $\{$*sun,*$\cdots$*,sky,*$\cdots\}$, $\{$*sky,*$\cdots$*,sun*$\}$. Therefore, alignment structure is not very useful here.

The HMM Model [12] assumes that there is a dependency between the a_j and a_{j-1} by making the alignment probability $P(\mathbf{a}|m,\mathbf{e}) = \prod_{j=1}^{m} p(a_j|a_{j-1},l,m)$ dependent on "$a_j - a_{j-1}$" instead of the absolute positions a_j. In our scenario this means that the knowledge of the previous alignment between a concept and a visterm can better predict the next possible alignment. Intuitively this idea should work in our context, when we align *sun* to a block and subsequently when trying to align *sky*, previous alignments can easily determine the most likely blocks to align to sky (the sky is always not far away from the sun). However the training procedure of the model requires the image to be flattened as a sequence of visterms (enumerate the block left to right and top to bottom), so that the adjacent blocks do not preserve this property. With this image representation the HMM model is able to capture only dependent alignments in the same row.

For Model 3,4,5 the translation probability is the following:

$$P(f|e) = \sum_{\mathbf{a}} P(\mathbf{f},\mathbf{a}|\mathbf{e}) = \sum_{\tau,\pi\in\langle\mathbf{f},\mathbf{a}\rangle} P(\varPhi|\mathbf{e})P(\tau|\mathbf{\Phi},\mathbf{e})P(\pi|\tau,\mathbf{\Phi},\mathbf{e}) \qquad (4)$$

where $P(\varPhi|\mathbf{e})$ represents the fertility probability and $P(\pi|\tau,\mathbf{\Phi},\mathbf{e})$ is the distortion probability.

The concept of *distortion* is useful when translating between languages with different word orders: English is a SVO language where Arabic is a VSO language (verb subject object). In order to use these models successfully, our training data should suffer from the same problem. Even though the *visterm* language has a structure, this one is lost when moving from a two-dimensions representation to a one-dimension representation. The *concept* language lack of structure, the concepts are enumerated as the annotators decided, each one with a different style. The same images can get either be annotated with different concepts or the concepts can be presented in different orders.

The other notion used on these advanced models is *fertility.* The fertility parameter gives for each English word how many French words it usually generates. For example in [10] the authors observed that the most likely fertility of *farmers* is 2 because it is most often translated as two words: *les agriculteurs*. We refer to fertility as the number of concepts associated with a block. In our data there is no such fixed number, if we have two images annotated with *house, tree, ...,* in one the *house* can occupy one block by itself and in another *house,tree* can be together. Depending on the resolution of the image, one block can be associated with either one or multiple concepts. Where in language the fertility of each words can be almost deterministically determined, it is not the same with visterms. As we can see neither distortion or fertility as stated offer additional information, instead they only add noise to the parameter estimation.

6 Image Annotation using Cross-Lingual Information Retrieval

The image annotation problem can alternatively be viewed as the problem of Cross-Lingual Information Retrieval (CLIR). In CLIR we have queries in a language "A" and the document collection in a language "B". The goal is to find the most relevant documents in language B for each query Q from language A. If we assume that language A is the language of concepts and B is the language of visterms, the task of image annotation becomes a CLIR problem. Suppose we would like to find for the concept c the most relevant images in our collection, we would rank each document using the following equation [15]:

$$p(c|d_V) = \alpha\big(\sum_{v \in d_V} p(c|v)p(v|d_V)\big) + (1-\alpha)p(c|G_C), \tag{5}$$

where c is a concept and d_V is a image document. Since the term $p(c|G_C)$ is the unigram probability of the concept c estimated on training data and does not depend on d_V, it will be dropped and the above formula can be rewritten as:

$$p(c|d_V) = \sum_{v \in d_v} p(c|v)p(v|d_V). \tag{6}$$

In order to compute $p(c|d_V)$ we need to estimate $p(v|d_V)$ and $p(c|v)$. The probability $p(v|d_V)$ is computed directly from the document d_V. The probability $p(c|v)$ is the probability of the concept c given that the visterm v is the document d_V; this is obtained as the translation probability estimated in the machine translation approach. As already mentioned each document is represented by a fixed number(105) of visterms. The visterm vocabulary is of size 3000. For most images $p(v|d_V)$ usually turns out to be close to $\frac{1}{105}$ for each visterm $v \in d_V$, i.e. the v's are unique.

However individual images are not able to produce a good estimate of $p(v|d_V)$. So we choose to estimate the prior probability over the training collection in the following ways:

$$TF_{Train}(v) = \frac{\#\text{ of } v \text{ in the collection}}{\#\text{ of visterms in the collection}}$$

$$DF_{Train}(v) = \frac{\#\text{ of documents with } v}{\#\text{ of documents in the collection}}$$

Since document frequency (DF) outperforms the term frequency (TF), $DF_{Train}(v)$ was used as a estimate of $p(v)$. Using $p(v)$ to approximate $p(v|d_V)$ and restricting the sum over only the visterms in the given document, we now have a score that is not a probability:

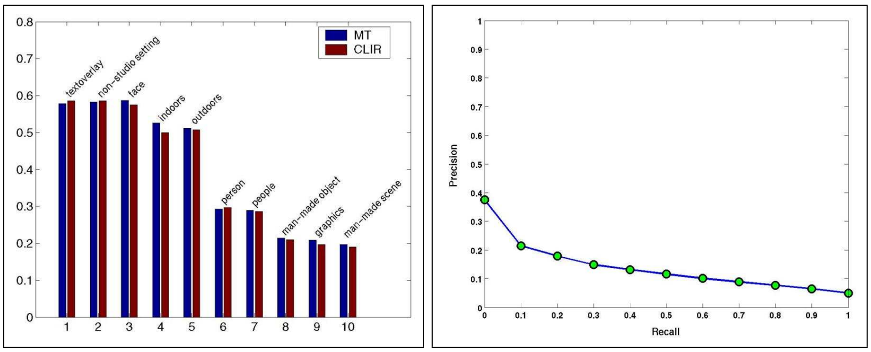
$$score(c|d_V) = \sum_{v \in d_v} p(c|v)DF_{Train}(v) \tag{7}$$

The annotation performance of the CLIR approach is shown in Table 3, the CLIR approach performs significantly better than our baseline Model 1 (p=0.04).

Figure 4-a compares the basic machine translation based approach with CLIR based approach using average precision values for the top 10 words. The recall-precision performance for CLIR is given in Figure 4-b.

Table 3. Annotation performance of CLIR approach.

System	mAP
Model1	0.124
CLIR	0.126



(a) (b)

Fig. 4. (a)Average Precision comparison between MT and CLIR based models for the top 10 concepts (b) Recall Precision performance for the CLIR annotation mode.

7 Discussion and future work

We conclude that the SMT (Statistical Machine Translation) approach [10] [12] to Image/Video retrieval is not tailored to this task and instead, we should look for newer approaches to “translation” models in this scenario. The IBM and HMM based text translation models have been developed to model the dependencies present in the translation of natural languages. However when applied to our task of image/video annotation, these powerful models are unable to improve modeling. This is mainly because our data - visterms and concept pairs - do not contain the same structure present in language pairs. Therefore additionally modeling power of the SMT model does not improve the ability of the model to predict new data. In contrast simpler translation models such as IBM-1 which do not rely much on the structure of the language pairs perform better when applied to the annotation task. We also note that the IBM models were originally designed to deal with languages that generate one-dimensional strings, in our task the *visterm* language generates two-dimensional strings and the *concept* language generates string without any particular order. As already seen in the MT community, the IBM models are not the only solution to the problem. Researchers are developing translation systems using *syntactic and parsing knowledge*, [13] [14]. Along these lines we should start to develop new translation systems that suit our data best.

8 Acknowledgements

This research is partially supported by ONR award number N000140110685, National Science Fundation grant number ITR-IIS-0121285, and by TÜBİTAK Career grant number 104E065 and grant number 104E077.

References

1. O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In The Fifteenth International Conference on Machine Learning, 1998.
2. L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semi-automatic image annotation. In INTERACT2001, 8th IFIP TC.13 Conference on Human-Computer Interaction, Tokyo, Japan July 9-13, 2001.
3. J. Li and J.Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. on Pattern Analysis and Machine Intelligence, 25(10):14, 2003.
4. Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In First International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.
5. K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. In Int. Conf. on Computer Vision, pages 408415, 2001.
6. J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In 26th Annual International ACM SIGIR Conference, July 28-August 1, 2003, Toronto, Canada.
7. D.M. Blei and M. I. Jordan. Modeling annotated data. In 26th Annual International ACM SIGIR Conference, July 28-August 1, 2003, Toronto, Canada.
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 Seventh European Conference on Computer Vision (ECCV), volume 4, pages 97112, Copenhagen, Denmark, May 27 - June 2 2002.
9. Iyengar G. et.al. "Joint Visual-Text Modeling", CLSP Workshop 2004, Johns Hopkins University, July-August 2004.
10. P.F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263311, 1993.
11. TREC Video Retrieval Evaluation http://www-nlpir.nist.gov/projects/trecvid/
12. S. Vogel, H. Ney, and C. Tillmann. 1996. "HMM Based Word Alignment in Statistical Translation." In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96),* pp. 836-841, Copenhagen, Denmark, August.
13. Kenji Yamada and Kevin Knight, "A Syntax-based Statistical Translation Model", in *Meeting of the Association for Computational Linguistics,*", pp. 523-430, 2001
14. I. Dan Melamed. "Statistical Machine Translation by Parsing", in *Proceedings of the 42nd Annual Conference of the Association for Computational Linguistics (ACL-04),* Barcelona, Spain.
15. Xu et. al., "Evaluating a probabilistic model for cross-lingual information retrieval" in *Proceedings of the 24th annual international ACM SIGIR,* New Orleans, Louisiana, United States. 2001.
16. Giza++ Toolkit, http://www.fjoch.com/GIZA++.html