Appeared in IEEE ICDE 2009 Conference.

Using Semantics for Speech Annotation of Images

Chaitanya Desai Dmitri V. Kalashnikov Sharad Mehrotra Nalini Venkatasubramanian

Computer Science Department
University of California, Irvine

I. Introduction

Digital cameras and multimedia capture devices are becoming increasingly popular to take pictures. Annotating these pictures is important to support their browsing and retrieval. Fully automatic image annotation techniques typically rely entirely on visual properties of the image. The state of the art image annotation systems of this kind work well in detecting generic object classes: car, horse, motorcycle, airplane, etc. However, certain characteristics of the image are hard to capture using strictly the visual properties. These include location (Paris, California, San Francisco, etc), event (birthday, wedding, graduation ceremony, etc), people (John, Jane, brother, etc) and abstract qualities referring to objects in the image (beautiful, funny, sweet, etc) among others. The more conventional method of annotation that relies completely on human input has several limitations as well. Typing tags using the keypads of such devices can be cumbersome and error-prone. Secondly, delay in tagging may result in a loss of context in which the picture was taken (e.g., user may not remember the names of the people/structures in the image).

This presents an opportunity for using speech as a modality to annotate images and/or other multimedia content. Most camera devices have a built-in microphone. In principle, some of the challenges associated with both, fully automatic annotation as well as manual tagging can be alleviated if the user were to use speech as a medium of annotation. Ideally, the user would take a picture and speak the desired tags into the device's microphone. A speech recognizer would transcribe the audio signal into text. The speech to text transcription can happen either on the device itself or be done on a remote machine. The transcribed text can be used as tags for the image, exactly as the user intended. One of the biggest bottlenecks facing such systems is the accuracy of the underlying speech recognizer. Even speaker dependent recognition systems can make mistakes in noisy environments. If the recognizer's output is considered as is for annotation, then poor recognition will lead to poor quality tags. Our work tries to address this issue by incorporating outside semantic knowledge to improve interpretation of the recognizer's output, as opposed to blindly believing what the recognizer suggests. To improve interpretation of speech output, we exploit the fact that most speech recognizers provide alternate hypotheses for each utterance. The main contribution of this paper is our approach for annotating images using speech as the input modality. The approach employs a probabilistic model for computing the joint probability of a given combination of tags using a Maximum Entropy solution. The extensive empirical evaluation demonstrates the advantage of the proposed solution, that leads to a significant improvement of quality of speech annotation.

This research was supported by NSF Awards 0331707 and 0331690, and DHS Award EMW-2007-FP-02535.

II. Problem Definition

We consider a setting wherein the user intends to annotate an image with a sequence $G = (g_1, g_2, \ldots, g_K)$ of K "ground truth" tags. Each tag g_i can be either a single word or a short phrase of multiple words, such as Niagara Falls, Golden Gate Bridge, and so on. Since a tag is typically a single word, we will use 'tag' and 'word' interchangeably.

A. N-Best Lists

To accomplish the annotation task, the user speaks out each of the words g_i for $i = 1, 2, \ldots, K$. These K words are then processed by a speech recognizer. We assume that the recognizer is trained to recognize a delimiter between each of these K utterances. The recognizer's task is to correctly recognize these words so that they can then be assigned as tags to the image. However, noisy environments and unrestricted vocabularies can increase the recognizer's uncertainty in its hypotheses. The recognizer might propose several *alternatives* for each utterance of a word. Thus, the output of the recognizer is a sequence $\mathcal{L} = (L_1, L_2, \ldots, L_K)$ of K N-best lists for the K utterances.

Each N-best list $L_i = (w_{i1}, w_{i2}, \ldots, w_{iN})$ consists of N words that correspond to the recognizer's alternatives for word g_i. Observe that list L_i might not contain the ground truth word g_i. The words in a N-best lists L_i are typically output in a ranked order. Thus, when the recognizer has to commit to a single word for each utterance, it would set $N = 1$ and output $(w_{11}, w_{21}, \ldots, w_{K1})$ as its answer. While w_{i1} has the highest chance of being the correct word, in practice it may be the incorrect option. This presents the need for an approach that can smartly disambiguate between the alternatives.

B. Answer Quality

We noted earlier that each L_i may or may not contain the ground truth. Let us define a *sequence* as a K-dimensional vector $W = (w_1, w_2, \ldots, w_K)$, where each w_i is either an element from list L_i or is equal to `null`, where $w_i = \texttt{null}$ encodes the fact that the algorithm believes that the list L_i does

L_1	L_2	L_3	L_4	L_5
w_{11}=pain	w_{21}=prose	w_{31}=**garden**	w_{41}=**flower**	w_{51}=sad
w_{12}=**Jane**	w_{22}=nose	w_{32}=harden	w_{42}=power	w_{52}=wad
w_{13}=lane	w_{23}=**rose**	w_{33}=jordan	w_{43}=shower	w_{53}=bad
w_{14}=game	w_{24}=crows	w_{34}=pardon	w_{44}=tower	w_{54}=dad

TABLE I
SAMPLE N-BEST LISTS $\mathcal{L} = (L_1, L_2, L_3, L_4, L_5)$.

not contain g_i. Now we can define the *quality* of a sequence $A = (w_1, w_2, \ldots, w_K)$ by adapting the standard IR metrics of precision, recall, and F-measure [1]. Namely, if $|A| = 0$ then $Precision(A) = Recall(A) = 0$. If $|A| > 0$ then $Precision(A) = \frac{|A \cap G|}{|A|}$ and $Recall(A) = \frac{|A \cap G|}{|G|} = \frac{|A \cap G|}{K}$, where $|A \cap G|$ is the number of w_i such that $w_i = g_i$. The F-measure is computed as the harmonic mean of the precision and recall. Thus, our goal can be viewed as that of designing an algorithm that produces high quality answer for any given $\mathcal{L}$.

Having defined the quality of an answer, we can make several observations. First, for a given $\mathcal{L}$ the best answer sequence is the sequence $A = (w_1, w_2, \ldots, w_K)$ such that $w_i = g_i$ if $g_i \in L_i$ and $w_i = \texttt{null}$ if $g_i \notin L_i$. Another related observation is that there is a theoretic upper bound on the achievable quality of any sequence A for a given $\mathcal{L}$. Specifically, assume that only M out of K N-best lists contain the ground truth tags, where $M \leq K$. Then the maximum reachable value of $|A \cap G|$ is M. Thus, if $M = 0$ then for any answer A it follows that $Precision(A) = Recall(A) = 0$. If $M > 0$ then the maximum reachable precision is $\frac{M}{M} = 1$ and maximum recall is $\frac{M}{K}$ which is less than 1 when $M < K$.

C. Overall Algorithm

We will consider the set $\mathcal{W}_\mathcal{L} = \{W\}$ of all N^K possible answer sequences given $\mathcal{L}$. On each such sequence $W \in \mathcal{W}_\mathcal{L}$ a score $S(W)$ will be assigned based on the joint distribution of the tags in W. The algorithm chooses a sequence W^* as its final answer by selecting among all $W \in \mathcal{W}_\mathcal{L}$ a sequence with the maximum overall score $W^* = \text{argmax}_{W \in \mathcal{W}_\mathcal{L}} S(W)$ and then applying a null detection procedure to W^* to compute the final answer.

D. Notational Example

As an example, suppose that the user takes a picture of her friends *Jane* in a garden full of roses, and provides the utterances with $K = 5$ words: $G = (g_1 = \textit{Jane}, g_2 = \textit{rose}, g_3 = \textit{garden}, g_4 = \textit{flower}, g_5 = \textit{red})$. Then, the corresponding set of five N-best lists for $N = 4$ could be as illustrated in Table I. If the recognizer has to commit to a single word per utterance, its output would be $(pain, prose, garden, flower, sad)$. That is, only 'garden' and 'flower' would be chosen correctly. This motivates the need for an approach that can disambiguate between the different alternatives in the list. Theoretically, the best possible answer would be $(Jane, rose, garden, flower, \texttt{null})$. The last word is null since list L_5 does not contain the ground truth tag $g_5 = red$. Therefore the maximum achievable precision is 1 and recall is $\frac{4}{5}$. Suppose some approach is applied to this case, and its answer is $A = (Jane, rose, garden, power, \texttt{null})$, that is, it picked 'power' instead of 'flower' and thus only 'Jane', 'rose', and 'garden' tags are correct. Then $Precision(A) = \frac{3}{4}$ and $Recall(A) = \frac{3}{5}$.

III. USING SEMANTICS FOR DENSITY ESTIMATION

Here we show how we compute the score of a sequence $W = (w_1, w_2, \ldots, w_K)$ as the joint probability $P(w_1, w_2, \ldots, w_K)$ for an image to be annotated with tags $w_1, w_2, \ldots, w_K$ using the approach of Maximum Entropy (ME). This probability is inferred based on how a collection of images has been annotated in the past. The ME approach reduces the problem of computing $P(w_1, w_2, \ldots, w_K)$ to a *constrained optimization problem*. It allows us to compute joint probability $P(w_1, w_2, \ldots, w_K)$ based on only the values of known correlations in data. The approach hinges on the information theoretic notion of entropy [6]. For a probability distribution $P = (p_1, p_2, \ldots, p_n)$, where $\sum p_i = 1$, the entropy $H(P)$ is computed as $H(P) = -\sum_{i=1}^{n} p_i \log p_i$ and measures the uncertainty associated with P. Entropy $H(P)$ reaches its minimal value of zero in the most certain case where $p_i = 1$ for some i and $p_j = 0$ for all $j \neq i$. It reaches its maximal value in the most uncertain uniform case where $p_i = \frac{1}{n}$ for $i = 1, 2, \ldots, n$. We will use a support-based method to decide whether the probability can be estimated directly from data. Specifically, if $K = 1$, or if $K \geq 2$ and $n(w_1, w_2, \ldots, w_K) \geq k$, where k is a positive integer value of *support*, then there is sufficient support to estimate the joint probability directly from data and $P(w_1, w_2, \ldots, w_K)$ is computed using a frequency based maximum likelihood estimate along with Lidstone's estimation that assumes a uniform prior on unseen sequences [2]–[4]. In particular, a support-based estimate would be $P(w_1, w_2, \ldots, w_K) = \frac{n(w_1, w_2, \ldots, w_K) + \lambda}{N_I + \lambda |V|}$. We will refer to such $P(w_1, w_2, \ldots, w_K)$ as *known* probabilities. Cases of $P(w_1, w_2, \ldots, w_K)$ where $K \geq 2$ but $n(w_1, w_2, \ldots, w_K) < k$ do not have sufficient support. They will be handled by the ME approach. We will refer to them as *unknown* probabilities.

To compute $P(w_1, w_2, \ldots, w_K)$ the ME approach considers the power set $\mathcal{S}$ of set $\{w_1, w_2, \ldots, w_K\}$, that is, the set of all its subsets. For instance, the power set of $\{w_1, w_2, w_3\}$ is $\{\{\}, \{w_1\}, \{w_2\}, \{w_1, w_2\}, \{w_2, w_3\}, \{w_1, w_2, w_3\}\}$. We can observe that for some of the subsets $S \in \mathcal{S}$ the probability $P(S)$ will be known and for some it will be unknown. Let $\mathcal{T}$ be the *truth* set, i.e., the set of subsets for which $P(S)$ is known: $\mathcal{T} = \{S \in \mathcal{S} : P(S) \text{ is known}\}$. The values of $P(S)$, where $S \in \mathcal{T}$, will be used to define the constraints for the constrained optimization problem.

To compute $P(w_1, w_2, \ldots, w_K)$ the algorithm considers *atomic annotation descriptions*, which are tuples of length K, where the i-th element can be only either w_i or $\overline{w}_i$. Here w_i means tag w_i is present in annotations and $\overline{w}_i$ means w_i is absent from them. For instance, description $(w_1, w_2, \overline{w}_3)$ refers

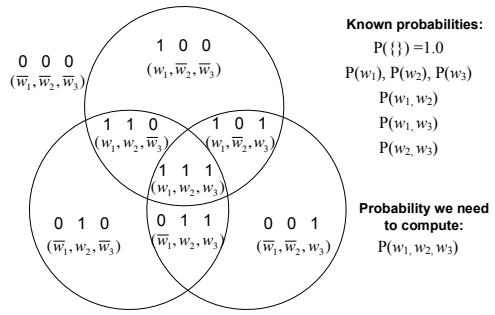

Fig. 1. Probability Space.

to all image annotations where tags w_1 and w_2 are present and w_3 is absent. Each such description can be encoded with a help of a bit string b, where 1 corresponds to w_i and 0 to $\overline{w}_i$. For instance $(w_1, w_2, \overline{w}_3)$ can be encoded as $b = 110$. Let A_S be the atom set for S, defined as the set of all possible bit strings of size K such that for each $b \in A_S$ it holds that if $w_i \in S$ then $b[i] = 1$, for $i = 1, 2, \ldots, K$. For instance for $K = 3$ and $S = \{w_1, w_2\}$ set $A_S = \{110, 111\}$, whereas for $K = 3$ and $S = \{w_2\}$ set $A_S = \{010, 011, 110, 111\}$.

Let x_b denote the probability to observe an image annotated with the tags that correspond to bit string b of length K. Figure 1 illustrates the probability space with respect to all x_b for the case where $K = 3$. Then in the context of ME approach our goal of determining $P(w_1, w_2, \ldots, w_K)$ reduces to solving the following constrained optimization problem:

$$\begin{cases} \text{Maximize } Z = -\sum_{b \in B} x_b \log x_b \\ \text{subject to} \\ \quad \sum_{b \in A_S} x_b = P(S) \text{ for all } S \in \mathcal{T} \\ \text{and} \\ \quad x_b \geq 0 \text{ for all } b \end{cases} \tag{1}$$

Solving it will give us the desired $P(w_1, w_2, \ldots, w_K)$ which corresponds to $x_{11\cdots1}$. The constrained optimization problem can be solved efficiently by the method of Lagrange multipliers to obtain a system of optimality equations. Since the entropy function is concave, the optimization problem has a unique solution [5]. We employ the variant of the iterative scaling algorithm used by [4] to solve the resulting system.

A. Correlations

In this section we define the notion of the correlation $c(w_i, w_j)$ between any pair of words w_i and w_j. This will allow us to create a method for detecting `null` cases. Let the correlation $c(w_i, w_j)$ between two words w_i and w_j be defined as the Jaccard similarity:

$$c(w_i, w_j) = \begin{cases} \frac{n(w_i, w_j)}{n(w_i) + n(w_j) - n(w_i, w_j)} & \text{if } n(w_i, w_j) > 0; \\ 0 & \text{if } n(w_i, w_j) = 0. \end{cases} \tag{2}$$

Here, $n(w_i, w_j)$, $n(w_i)$, and $n(w_j)$ are the number of images whose annotation include, respectively, both the tags w_i and w_j, tag w_i, and tag w_j. The value $c(w_i, w_j)$ is always in $[0, 1]$ interval.

We can extend the notion of direct correlations to that of indirect correlations. Observe that even when two words may never have co-occurred together in any image, they could still be correlated to each other through other words. For instance, the words *beach* and *ocean* may be indirectly correlated through the word *sand*.

Suppose we start with a *base correlation graph* $\mathcal{G} = (V, E)$ whose nodes are tags in the vocabulary V. An edge is created per each pair of nodes w_i and w_j and labelled with the value of $c(w_i, w_j)$. The *base correlation matrix* $B = B_1$ of $\mathcal{G}$ is a $V \times V$ matrix with elements $B_{ij} = c(w_i, w_j)$. Let P^2_{ij} be the set of all paths of length two in graph $\mathcal{G}$ from w_i to w_j. Then the indirect correlation $c_2(w_i, w_j)$ of length two for w_i and w_j is defined as the sum of contribution of each path $(x_0 x_1 x_2) \in P^2_{ij}$, where the contribution of each path is computed as the product of base similarities on its edges:

$$c_2(w_i, w_j) = \sum_{(x_0 x_1 x_2) \in P^2_{ij}} \prod_{i=1}^{2} c(x_{i-1}, x_i). \tag{3}$$

It can be shown that the corresponding similarity matrix B_2 can be computed as $B_2 = B^2$. The idea can be extended further by considering $c_k(w_i, w_j)$ and demonstrating that $B_k = B^k$. A similarity matrix A, that takes into account indirect similarities for $k = 1, 2, \ldots, m$ can be computed in a manner similar to that of diffusion kernels [7]. For instance, in the spirit of exponential diffusion kernels, A can be computed as $A = \sum_{k=0}^{m} \frac{1}{k!} \lambda^k B^k$, or, as $A = \sum_{k=0}^{m} \lambda^k B^k$. A and B_k for $k = 1, 2, \ldots, m$ are performed off line before processing of image annotations starts. Therefore very fast computation of A is not critical.

B. Detecting Nulls

This section discusses how A can be utilized for detecting null candidates. That is, detecting the situation that a given N-best list L_i is unlikely to contain the ground truth tag g_i.

First, we extend the notion of a base correlation graph $\mathcal{G}$ to that of indirect correlation graph $\mathcal{G}_{ind}$. Like in $\mathcal{G}$, the nodes of $\mathcal{G}_{ind}$ are the tags $w_i \in V$, but each edge (w_i, w_j) is now labelled with the value of A_{ij}.

Let $W^* = (w_1, w_2, \ldots, w_K)$ be the sequence with the highest score among all the possible N^K sequences for a given sequence of N-best lists $\mathcal{L}$. If list $L_i \in \mathcal{L}$ does not contain the ground truth tag g_i, then $w_i \neq g_i$. We can observe that when such situations occur, it is likely that w_i will not be strongly correlated with the rest of the tags in W^*. We can now design the null detection procedure. It takes $W^* = (w_1, w_2, \ldots, w_K)$ as input and analyzes each $w_i \in W^*$. If $A(w_i, w_j) < \tau$ for $j = 1, 2, \ldots, K$, $j \neq i$, and a threshold value τ, then w_i is considered to be *isolated* in $\mathcal{G}_{ind}$, in terms of correlations, from the rest of the tags. Such isolated tags are substituted with null values.

IV. Experiments

Dataset:. Our set of images was obtained by crawling a popular image hosting website, namely Flickr. We start off

by downloading 60000 Flickr images with their ground truth annotations. We randomly set aside 20% of the data for testing (will be called D_{test}) and 80% for training (D_{train}). We will use portions of D_{test} for testing. The size of the vocabulary is $|V| = 18285$.

We randomly picked 102 images from D_{test} and annotated them (generating the N-best lists) using a popular commercial off-the-shelf recognizer *Dragon v.8*. We will call this annotated set $\bar{D}_{test}$. The annotations were performed in a Low noise level. Low noise level corresponds to a quiet university lab environment. All non-English words were removed before using *Dragon* to create these N best lists.

Approaches. We will compare the results of three approaches:

- `Baseline` is the output of the recognizer (Dragon v.8).
- `ME` is the output of the proposed approach.
- `Upper Bound` is the theoretic upper bound achievable, see Section II. It depends on how many N-best lists contain the ground truth tag.

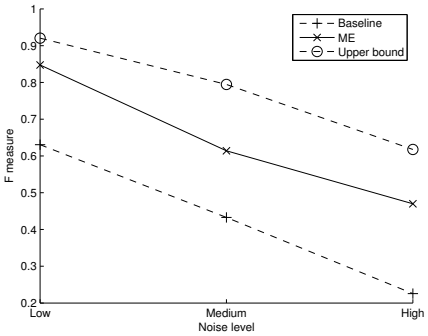

Fig. 2. F-measure vs. Noise.

Experiment 1. (Quality for Various Noise Levels.) We randomly picked 20 images from $\bar{D}_{test}$ and re-annotated them (i.e. created N best lists for ground truth tags) using *Dragon* in two additional noise levels: Medium and High. Medium and High levels were produced by introducing white Gaussian noise through a speaker.[1] Figure 2 shows the F-measure of the three approaches for the Low, Medium, and High noise levels on these 20 images. Since we had created $\bar{D}_{test}$ in a Low noise level on 102 images, for a fair comparison, the points corresponding to Low noise levels in the plots are averages over these 20 images, as opposed to the full 102 images. As anticipated, higher noise levels negatively affect performance of all three approaches. The performance of `ME` dominates `Baseline` performance for all the different noise levels.

When we considered images that are annotated with exactly K tags, we found the performance of `ME` is consistent across different values of K. For instance, for Low noise level and for $K = 2, 3, \ldots$, the F-measure of `ME` was consistently within 12% of F-measure for the `Upper Bound`. In addition, it was constantly better than the F-measure of `Baseline` by at least 15% for all K.

In the subsequent discussion we will refer to $\bar{D}_{test}$ data with the Low level of noise as just $\bar{D}_{test}$.

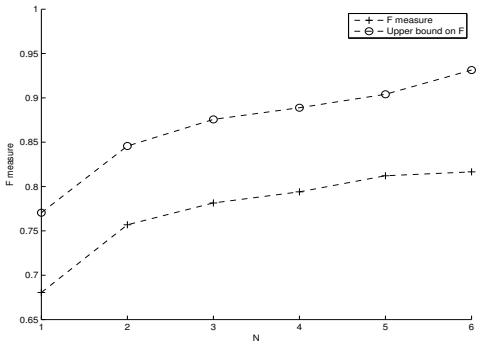

Fig. 3. F vs. N (Size of N-Best Lists)

Experiment 2. (Quality vs Size of N-Best Lists.) Figure 3 illustrate the F-measure as a function of the size of N-best list N on $\bar{D}_{test}$ data. For a given N, the N-best lists are generated by taking the original N-best lists from $\bar{D}_{test}$ data and keeping at most N first elements in them. Increasing N presents a tradeoff. As N increases, the greater is the chance that the ground truth element would appear in the list. At the same time, `ME` algorithm is faced with more uncertainty as there are more options to disambiguate between. The results demonstrate that the potential benefit from the former outweighs the potential loss due to the latter, as F-measure increases with N.

V. Conclusion:

In this paper, we have postulated the problem of using discrete speech utterances to annotate an image as that of disambiguation across multiple N-best lists. Our solution is based on the Maximum Entropy approach and uses correlations between tags in an existing corpus of images to set up the constrains of the corresponding constrained optimization problem. Our experiments suggest that the proposed approach gives a significant improvement in quality as compared to an approach that considers the best answer suggested by a popular off-the-shelf recognizer.

References

[1] R. Bayeza-Yates and B. Riberto-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] D. Jurafsky and J. Martin. *Speech and Language Processing*. Prentice-Hall, 2000.

[3] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[4] V. Markl, P. J. Haas, M. Kutsch, N. Megiddo, U. Srivastava, and T. M. Tran. Consistent selectivity estimation via maximum entropy. *VLDB J.*, 16(1):55–76, 2007.

[5] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty. Inducing features of random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(4), 1997.

[6] C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.

[7] J. Shawe-Taylor and N. Cristianni. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[1]To give a sense of the level of noise, High was a little louder than the typical volume of TV in a living room.