Disambiguation of Annotated Text of Audio Using Ontologies

Latifur Khan and Dennis McLeod
Department of Computer Science and
Integrated Media Systems Center
University of Southern California
Los Angeles, California 90089
[latifurk, mcleod]@usc.edu

ABSTRACT

To improve the accuracy in terms of precision and recall of an audio information retrieval system we have created a domain-specific ontology. Taking into account the shortcomings of keyword-based techniques, we have opted to employ a concept-based technique utilizing ontology. The key problem is the identification of appropriate concepts within annotated audio text. In the case of the association of irrelevant concepts with audio objects there is a loss of precision. On the other hand, if relevant concepts are discarded, a loss of recall will ensue. Therefore, we have proposed a novel automatic disambiguation algorithm which prunes as many irrelevant concepts as possible while at the same time retaining the largest possible number of concepts which are relevant. Through the use of techniques for the association of concepts in the ontology, we have devised a method for determining the correlation of concepts and are thus able to employ a correlation factor as a basis for the selection of concepts for audio objects. In trial implementations of our algorithm we have achieved a level of accuracy at which up to 76.9% of the objects identified in audio text have been associated with relevant concepts.

Keywords

Metadata, Ontology, Audio, SQL, Precision, Recall

1. INTRODUCTION

The development of technology in the field of digital media generates huge amounts of non-textual information, such as audio, video, and images, as well the more familiar textual information [16]. In general, the amount of information available via electronic means can easily overwhelm end-users. Further, any transfer of irrelevant information over the network to end-users wastes network bandwidth. Therefore, the need for user-customized information selection is clear. Among the non-textual media, audio is one of the most powerful and expressive. It is of special note that audio information can be of particular benefit to a person who is visually impaired, and for the general population, audio, as a streaming medium i.e. temporally extended, is an increasingly popular medium for capturing and presenting information. At the same time, its very properties as a medium, along with audio's opaque relationship to computers, present distinct technical problems from the perspective of data management.

The effective selection/retrieval of audio information entails several tasks, such as metadata generation (description of audio), and consequent selection of audio information in response to a query. One of the key challenges is finding ways to facilitate the provision of these responses through metadata. Since the goal in customized selection and delivery is high precision (little irrelevant information) and high recall (no omission of relevant information), we propose addressing this problem through the use of an ontology-based model.

Thus in this paper, we advance a model using a domain dependent ontology. An ontology is a collection of concepts and their inter-relationships that collectively provide an abstract view of an application domain. Relevant to our purpose, ontologies can be employed to facilitate metadata generation and information selection requests. These metadata can be created by following current state-of-the-art procedures in speech recognition technology. This will involve the use of a fully automated content extraction [10] technique (speech to text conversion), and selected content extraction using word-spotting [24], which determines the occurrence of keywords in audio where these keywords are derived from ontologies. It is in view of the shortcomings of keyword-based technique [23] that we adopt a concept-based technique for information selection requests using ontologies [1, 8].

The key problem in the use of this technique will be to identify and match appropriate concepts from the annotated text of audio on the one hand and user requests on the other. In this it is critically important to make sure that irrelevant concepts will not be associated and matched, and that relevant concepts will not be discarded. In other words, it is important to insure that high precision and high recall will be preserved during user requests. At the outset of the process, multiple concepts may be selected. Some of these concepts will be relevant, others will be irrelevant. If irrelevant concepts are permitted to become metadata, precision will be hurt. Thus we need an automatic disambiguation algorithm which will prune irrelevant concepts, while allowing relevant concepts to become metadata.

Thus, in order to maximize precision and recall, we propose using a novel scalable automatic disambiguation algorithm. For automatic disambiguation within an ontology a set of *regions* representing different concepts can be defined. The concepts, as they appear in a given region, will be mutually disjoint from the concepts of other regions. This becomes the basis for determining a group of appropriate concepts for a given keyword or collection of keywords. In short, keywords are matched to the concepts of a given ontology, and then the region within the ontology in which the greatest number of selected concepts occurs is determined. This region, the one containing the maximum number of selected concepts, will then be used for annotation. The selected concepts of other, different regions will be automatically pruned.

Thus it can be seen that disambiguation approaches the determination of which concepts are relevant by matching keywords to the concepts of a given ontology, determining the region within the ontology in which the greatest number of selected concepts occurs, and then using that region for annotation. The selected concepts of other different regions will be automatically pruned. A simple example will make this clear. The keyword "Charlotte" for a particular audio is associated with two concepts of the ontology, "Charlotte Hornets" and "UNC Charlotte" respectively. One is in the region encompassing professional basketball, the other in the region encompassing college basketball. Thus, at various levels of complexity beyond this simplified example, the disambiguation technique used to distinguish between concepts is based on the general idea that any set of keywords occurring together in context will determine appropriate concepts for one another, i.e. fall into the same region, in spite of the fact that each individual keyword is multiply ambiguous. Any keyword alone will determine a group of concepts which are both relevant and irrelevant, and which can occur in different regions.

Next, we will need to have a way of dealing with the possibility that even within a region selected for annotation a given keyword will match more than one concept. In other words, multiple ambiguous concepts will have been selected for a particular keyword, necessitating further disambiguation. In order to further prune irrelevant concepts we will need to determine correlation between the concepts selected in a given region. When concepts are correlated, the scores of concepts strongly associated will be given greater weight in the final determination of concepts to be annotated. This association will be based on minimal distance in the ontology and the matching scores of concepts based on the number of keywords they match. Thus, selected concepts which correlate with each other will have a higher score, and a greater probability of being retained than non-correlated concepts. If scores of particular ambiguous concepts fall below a certain *threshold*, which will be a minimum score chosen for selected concepts for that particular audio, these concepts are pruned. It is important to note that in this paper we have not addressed any issues related to selection of audio information (see [1]).

At present, an experimental prototype for the implementation of the model is in development and under study. As of today, our working ontology has around 7,000 concepts for the sports news domain, with 15 hours of audio stored in the database. For sample audio content we use CNN broadcast sports and Fox Sports audio, along with closed captions. Using our disambiguation algorithm, these associated closed captions are connected with the ontology. The performance of our disambiguation algorithm has been studied by considering what percentage of objects associate with relevant concepts as well as the impact of threshold values on the problem of relevant association. We have observed that through the use of our disambiguation algorithm 90.5% of the objects successfully associate with concepts of ontologies, and 9.5% of the objects fail to associate with any concept of ontologies. Without the use of this pruning technique, only 60.8% of the objects are associated exclusively with relevant concepts, and 29.7% of the objects are associated with both relevant and irrelevant concepts. With the increasing threshold for inclusion we can demonstrate that fewer irrelevant concepts are retained. However, some relevant concepts may also be discarded. For this reason, it is important to choose a threshold value carefully. We have determined that the basis for selecting a threshold value depends entirely on the dataset. In our case, up to 76.9% of the objects are associated with relevant concepts.

The remainder of this paper is organized as follows. Section 2 covers some related work on disambiguation. Section 3 describes how we can use an ontology which will support metadata generation for audio. Section 4 discusses different mechanisms for annotating audio based on current state-of-the-art speech technology and presents our automatic disambiguation algorithm along with some refinements. Section 5 describes the current implementation status of our system and some results of the disambiguation algorithm. Finally, section 6 presents our conclusions and plans for future work.

2. Related Works

To place our research in the context of information retrieval effectiveness (precision and recall), we would like to summarize key related efforts. First, we will discuss a traditional keyword-based technique (vector space model) and next, a concept-based model using ontologies of the type we employ in order to overcome the shortcomings of a keyword-based search [1, 8].

In the case of a keyword based search (vector space model), documents are only retrieved in response to keywords specified by the user. Query expansion mechanisms can be employed as a way of addressing this limitation. Additional search terms are added to the original query, based on the statistical co-occurrence of terms. However, attempts at query expansion of this nature have not been very successful, since the use of a statistical method does not result in good control over which terms should be added and which terms pruned out. Although recall is expanded through the addition of new terms, this occurs at the expense of deteriorating precision [2, 7].

One way to demonstrate the superiority of the concept based technique over that of keyword search is to explore the treatment of a specific example. For example, when a query is specified in terms of motor vehicle, new terms: bus, truck, and car, are added to the original query. However, if the intent of the original query is to retrieve information about automobiles, the addition of the terms bus and truck is not helpful. In this situation, we require a conceptual hierarchy where one concept subsumes other concepts [8]. In the example given, the concept automobile would rest on the top of a hierarchy in which a variety of sub-concepts would be enumerated. In our model, this type of hierarchy constitutes an ontology within which concepts related to queries and

documents can be mapped into conceptual space in which measures of similarity are applied to these concepts.

Historically ontologies have been employed to achieve better precision and recall in the text retrieval system. Here, attempts have taken two directions, query expansion through the use of semantically related-terms and the use of conceptual distance measures, as in our model. Among attempts using semantically related terms, query expansion with a generic ontology, WordNet [19] has shown to be potentially relevant to enhanced recall, as it permits matching relevant documents to a query that do not contain any of the original query terms. Voorchees [26] manually expands 50 queries over aTREC-1 collection using WordNet, and observes that expansion was useful for short, incomplete queries, but not promising for complete topic statements. Further, for short queries, automatic expansion is not trivial; it may degrade rather than enhance retrieval performance. The notion of conceptual distance between query and document provides an approach to modeling relevance. Smeaton et al. [27] and Gonzalo et al. [21] focus on managing short and long documents, respectively. Note here that queries and document terms are manually disambiguated using WordNet. In our case, query expansion and disambiguation algorithms are fully automatic.

3. Metadata Acquisition

In this section, we describe how we can use ontology to facilitate metadata generation.

An ontology is a specification of an abstract, simplified view of the world that we wish to represent for some purpose [5, 9]. Therefore, an ontology defines a set of representational terms that we call *concepts*. Inter-relationships among these concepts describe a target world. An ontology can be constructed in two ways, domain dependent and generic. CYC [17], WordNet [19], or Sensus [20] are examples of generic ontologies. For our purposes, we choose a domain dependent ontology. First, this is because a domain dependent ontology provides concepts in a fine grain, while generic ontologies provide concepts in coarser grain. Second, a generic ontology provides a large number of concepts that may contribute large speech recognition error.

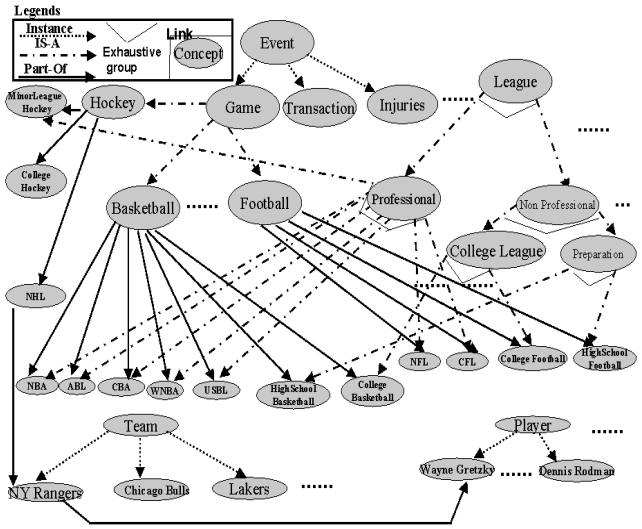


Figure 1. A Small Portion of an Ontology for Sports Domain

Figure 1 shows an example ontology for sports news. This ontology is usually obtained from generic sports terminology and domain experts [6]. This ontology is described by a directed acyclic graph (DAG). Here, each node in the DAG represents a concept. In general, each concept in the ontology contains a label name and a synonyms list. Note also that this label name is unique in the ontology. Further, this label name is used to serve as association of concepts with audio objects. The synonyms list of a concept contains vocabulary (a set of keywords) through which the concept can be matched with user requests. Formally, each concept has a synonyms list ($l_1, l_2, l_3, ..., l_i, ..., l_n$) where user requests are matched with this l_i what we call *element* of list. Note that a keyword may be shared by multiple concepts' synonyms lists. For example, player "Bryant Kobe," "Bryant Mark," "Reeves Bryant" share common word "Bryant" which may create ambiguity problem. Moreover, each of them belongs to league NBA. Hence, each of these concepts' label is prefixed with concept, NBA's label that allows to make efficient query generation for upper level concepts. The sample contents of concepts are as follows:

•NY Rangers:
–Label:- NHLTeam11
–IS-A:-
–Instance:- Team
–Part-of:NHL
–Synonyms list:- NY Rangers, New York Rangers, …

•NHL:
–Label:-NHL
–IS-A: -Professional
–Instance:-
–Part-of:-Hockey
–Synonyms list:-NHL, National Hockey League, ...

Thus, the labels for the concepts NY Rangers and NHL are NHLTeam11, and NHL respectively. The Concept NY Rangers is associated with concepts, Team and NHL, through Instance and Part-Of inter-relationships.

3.1 Inter-Relationships

In the ontology, concepts are interconnected by means of inter-relationships. If there is a inter-relationship R, between concepts C_i and C_j, then there is also a inter-relationship R′ between concepts C_j and C_i. In Figure 1, inter-relationships are represented by labeled arcs/links. Three kinds of inter-relationships are used to create our ontology: IS-A, Instance-Of, and Part-Of. These correspond to key abstraction primitives in object-based and semantic data models [3].

IS-A: This inter-relationship is used to represent concept inclusion. A concept represented by C_j is said to be a specialization of the concept represented by C_i if C_j is kind of C_i. For example, "NFL" is a kind of "Professional" league. In other words, "Professional" league is the generalization of "NFL." In Figure 1, the IS-A inter-relationship between C_i and C_j goes from generic concept C_i to specific concept, C_j represented by a broken line. The IS-A inter-relationship can be further categorized into two types: *exhaustive group* and *non-exhaustive group*. An exhaustive group consists of a number of IS-A inter-relationships between a generalized concept and a set of specialized concepts, and places the generalized concept into a categorical relation with a set of specialized concepts in such a way so that the union of these specialized concepts is equal to the generalized concept. For example, "Professional" relates to a set of concepts, "NBA", "ABL", "CBA", ..., by exhaustive group (denoted by caps in Figure 1). Further, when a generalized concept is associated with a set of specific concepts by only IS-A inter-relationships that fall into the exhaustive group, then this generalized concept will not participate in the annotation and SQL query generation explicitly. This is because this generalized concept is entirely partitioned into its specialized concepts through an exhaustive group. We call this generalized concept a *non participant concept (NPC)*. For example, in Figure 1 "Professional" concept is NPC. On the other hand, a non-exhaustive group consisting of a set of IS-A does not exhaustively categorize a generalized concept into a set of specialized concepts. In other words, the union of specialized concepts is not equal to the generalized concept.

Specialized concepts inherit all the properties of the more generic concept and add at least one property that distinguishes them from their generalizations. For example, "NBA" inherits the properties of its generalization, "Professional" but is distinguished from other leagues by the type of game, skill of participant, and so on.

Instance-Of: This is used to show membership. A C_j is a member of concept C_i. Then the inter-relationship between them corresponds to an Instance-Of denoted by a dotted line. Player, "Wayne Gretzky" is an instance of a concept, "Player." In general, all players and teams are instances of the concepts, "Player" and "Team" respectively.

Part-Of: A concept is represented by C_j is Part-Of a concept represented by C_i if C_i has a C_j (as a part) or C_j is a part of C_i. For example, the concept "NFL" is Part-Of "Football" concept and player, "Wayne Gretzky" is Part-Of "NY Rangers" concept.

When a number of concepts are associated with a parent concept through IS-A inter-relationship, it is important to note that these concepts are disjoint, and are referred to as concepts of a disjoint type. When, for example, the concepts "NBA", "CBA", or "NFL" are associated with the parent concept "Professional," through IS-A, they become disjoint concepts. Moreover, any given object's metadata cannot possess more than one such concept of the disjoint type. For example, when an object's metadata is the concept "NBA," it cannot be associated with another disjoint concept, such as "NFL." It is of note that the property of being disjoint helps to disambiguate concepts for keywords during annotation.

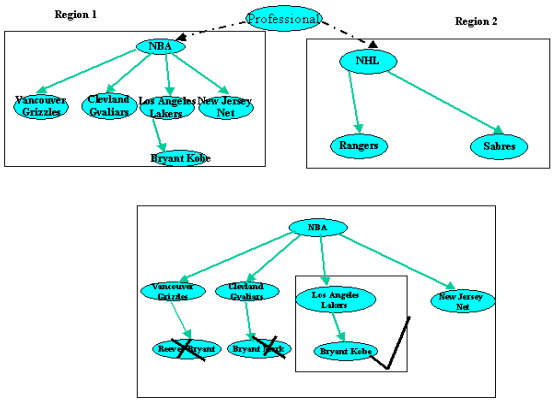


Figure 2. Different Regions of Ontology and Disambiguation of Concepts in a Region

Similarly, concept "college football", "college basketball" are disjoint concepts due to their associations with parent concept, "college league" through IS-A. Furthermore, "professional," and "non professional" are disjoint. Thus, we can say that "nba," "cba," "abl," "college basketball," and "college football," are disjoint. Each of these league and its team and player form a boundary what we call *region*. During annotation of concepts with an audio object we strive to choose a particular region. This is because an audio object can be associated with only one disjoint-type concept. However, it may be possible that a particular player may play in several leagues. In that case, we make multiple instances of the player. In other words, for each league he plays, we maintain a separate concept for him. This way we preserve disjoint-property.

Concepts are not disjoint, on the other hand, when they are associated with a parent concept through Instance-Of or Part-Of. In this case, some of these concepts may serve simultaneously as metadata for an audio object. An example would be the case in which the metadata of an audio object are team "NY Ranger" and player "Wayne Gretzky," where "Wayne Grezky" is Part-Of "NY Rangers."

4. Annotation

Annotation is the name for the process through which concepts are associated with audio objects. The completeness and accuracy of the annotation process contributes to the success of customization. In this section, we describe techniques for automatic and manual annotation. Using long pauses or speaker changes [4] after the segmentation of broadcast audio into what we call *audio objects*, we need to associate these objects with concepts of ontologies. Note that broadcast audio consists of multiple news items.

Automatic Annotation: Word-spotting techniques can provide the selected content extraction to make the annotation process automatic. Word-spotting, as noted, is a particular application of automatic speech recognition in which the vocabulary of interest is relatively small. Vocabularies of concepts from the ontology, excepting NPC concepts, can be used in our case. It is the job of

the recognizer to pick out only occurrences of keywords from this vocabulary in the speech (in our case audio object) to be recognized [14]. The output of a wordspotter is typically a list of keyword "hits" in this audio object. For example, if the occurrence of the concept, "NFL" is determined in a particular audio object, the object's metadata is the concept, "NFL." Fully automated content extraction may be employed [10].

Manual Annotation: Human intervention may be required to reduce speech recognition error. Furthermore, annotation can be provided in plain text, such as closed caption.

4.1 Disambiguation

We present an automatic disambiguation algorithm for choosing appropriate concepts for a group of keywords, and propose future further refinements.

For each audio object we need to find the most appropriate concept(s). Recall that using word-spotting or closed-captions we get a set of keywords which appear in a given audio object. It is possible that a particular keyword may be associated with more than one concept in the ontology. In other words, association between keyword and concept is one: many, rather than one: one. Therefore, disambiguation of concepts is required. The basic notion of disambiguation is that a set of keywords occurring together will determine the appropriate context for one another. To note our earlier example, base, bat, glove may have several interpretations as individual terms, but when taken together, the intent is obviously a reference to baseball.

For the disambiguation of concepts, we propose an efficient pruning algorithm based on the maximum-likelihood and correlation principle. Disambiguation algorithm first strives to disambiguate across several regions, and then disambiguates within a particular region. The basic procedure will be as follows:

First, we determine a region with the maximum number of selected concepts. Recall that a league and its team, and player form a region. Each concept in the ontology has a set of elements which constitute a synonyms list where each element has a set of keywords. For a concept to be selected, at least one keyword from the element must match with the annotated text of the audio object.

For example, the annotated text for a particular audio object might be: "*Lakers* keep grooving with 8th straight win. *Kobe Bryant* scores 21 points as the *Lakers* remain perfect on their *eastern* road trip with a 97-89 triumph over the *Nets. Bryant* discussed the eight game win streak and his performance in the All Star game."

The italic words are the keywords which are associated with the concepts of our ontology. The keywords "Lakers," and "Nets" are associated with the concepts "Los Angeles Lakers" and "New Jersey Nets" respectively. The keyword "Bryant" is associated with the concepts, "Reeves Bryant," "Bryant Mark," and "Bryant Kobe."

It is important to note that all these selected concepts are found in the region "NBA." The keyword "Eastern" is associated with the concepts "Eastern Washington," and "Eastern Michigan" which are in the region of "College Basketball." If we now choose only the concepts which appear in the NBA region, which is the region in which the greatest number of concepts occur, and maximize that number, the concepts "Eastern Washington" and "Eastern Michigan" will be eliminated, since they are not found in that region. Thus, we keep from among the concepts selected those which appear in NBA region, and prune other selected concepts.

It is important to note that in the selected region, in this case NBA, a keyword such as "Bryant" may be associated with more than one selected concept. This will necessitate further, disambiguation. We will want to know what other concept qualifies the concepts selected by keyword, "Bryant" through correlation. As noted above in the case of the keyword "Bryant," the concepts "Bryant Kobe," "Bryant Mark," and "Reeves Bryant" are selected. Among these ambiguous concepts, however only "Bryant Kobe" is correlated with another selected concept, in this case "Los Angeles Lakers." Therefore, "Bryant Kobe" is kept, and the concepts "Bryant Mark," and "Reeves Bryant" are thrown away (see Fig. 2).

Thus, we determine the correlation of selected concepts in the region in which the greatest number of keywords have been matched to the audio annotation, and within that region non-correlated ambiguous concepts are pruned. Finally, the selected concepts for this audio object are "New Jersey Nets," "Los Angeles Lakers," and "Bryant Kobe."

We have implemented the above idea using score-based techniques. To illustrate this technique we would first like to define some terms, and then present our score-based algorithm.

Formal Definitions:

Each selected concept contains a score, whether the concept is selected for partial or full match. Recall that in the ontology each concept (C_i) has a synonyms list ($l_1, l_2, l_3, ..., l_i, ..., l_n$). Keywords of annotated text are strived to match with each keyword on the element l_j of a concept and the score for l_j is calculated based on number of matched keywords of l_j. Maximum score of these scores is defined as score for this concept. Furthermore, when two concepts are correlated, scores are inversely affected by their position (distance) in the ontology.

Definition 1: Element-score (EScore): Element-score of an element l_j for a particular concept C_i is the number of keywords of l_j matched with keywords in the annotated text divided by total number of keywords in l_j.

$$EScore_{ij} \equiv \frac{\#\text{of keywords of lj matched}}{\#\text{ of keywords in } lj}$$

Normalization is used to nullify the effect of the length of l_j.

Definition 2: Concept-score (Score): Concept-score for a concept, C_i is the maximum score of all its element-scores.

Thus, $Score_i \equiv \max \text{EScore}_{ij}$ where $1 \le j \le n$

Definition 3: Semantic distance (SD (C_i , C_j)):between concepts C_i and C_j is defined as the shortest path between two concepts, C_i and C_j in the DAG. Note that if concepts are in the same level and no path exists, semantic distance is infinite. For example, semantic distance between concepts "NBA" and team "Lakers" is 1. This is because the two concepts are directly connected via Part-Of inter-relationship. Similarly, semantic distance between "NBA," and "Bryant Kobe" is 2. Semantic distance between "Los Angeles Lakers", and "Portland Blazers" is infinite.

Definition 4: Propagated-score (S_i): Propagated-score of a concept, C_i is defined by summing its concept-score plus other correlated concepts'(C_j) concept-scores divided by semantic distance between C_i and C_j.

Initially, $S_i = Score_i$

When C_i is correlated with C_j

$$S_i \equiv S_i + \frac{Score_j}{SD(C_i, C_j)}$$

$$S_j \equiv S_j + \frac{Score_i}{SD(C_i, C_j)}$$

Thus, when two concepts are correlated with each other where semantic distance is greater than one, they will have lower S_i and S_j as compared to the concepts of same concept-scores when their semantic distance is 1. This is because for higher semantic distance concepts are correlated in a broader sense. Thus, correlated concepts have higher S_i as compared to non-correlated concepts. The pseudo code for the disambiguation algorithm is as follows:

For each audio object
  Find concepts (C_1, C_2, C_3, …,C_i, …, C_m) that are associated with keywords of annotated text
 For each region, R
  $CScore_R = 0$
//Sum of all selected concepts concept-score for a region, R
For each keyword
  If selected ambiguous concepts (C_{k+1} ,C_{k+2}, …C_{k+r-2}, C_{k+r-1}, C_{k+r}) are in this region, R
  //Calculate their average concept-score, $CScore_{RA}$

$$CScore_{RA} \equiv \frac{Score_{ck+1} + Score_{ck+2} + ... + Score_{ck+r}}{r}$$

$$CScore_R \equiv CScore_R + CScore_{RA}$$

  Else
  If a non ambiguous concept C_1 is selected in this region, R

$$CScore_R \equiv CScore_{RA} + Score_{c1}$$

Choose a region with maximum score, $CScore_R$
and prune selected concepts in different regions

For the selected region
  Determine correlation of concepts (C_i, C_j, …) and update their propagated-scores by

$$S_i \equiv S_i + \frac{Score_j}{SD(C_i, C_j)}$$

$$S_j \equiv S_j + \frac{Score_i}{SD(C_i, C_j)}$$

//Prune non-correlated ambiguous concepts
  Determine maximum score S_{max} among all selected concepts' propagated score S_i for this object
For each ambiguous concept's propagated score S_i
  If ($S_i < S_{max}$ * threshold)
    Simply discard this concept which has S_i
  Else
    Keep this concept

Figure 3. Pseudo code for Disambiguation Algorithm

As noted above, there is a trade-off associated with the selection of value of threshold (γ); γ can be 0, 0.1, 0.2,… For high values of γ, we may loose some relevant concepts and at the same time discard many irrelevant concepts for audio objects. On the other hand, for a lower value of γ, we may keep many irrelevant concepts along with those which are correct. Our goal, for a given audio object, is to keep as many relevant concepts as possible and to throw away the maximum number of irrelevant concepts. By increasing γ, we may discard many ambiguous concepts. In this case, some of those discarded are indeed irrelevant for the object, and by throwing out these concepts better precision can be achieved. This is because in the latter case a given irrelevant object will not be retrieved when the user query is related to one of these discarded concepts.

For example, the concepts "Los Angeles Lakers," "New Jersey Nets," "Bryant Kobe," "Bryant Mark," and "Reeves Bryant," are selected in the selection of region "NBA." S_i, propagated-scores for these concepts are 1.5, 0.5, 1.5, 0.5, 0.5 respectively. Note that "Los Angeles Lakers," and "Bryant Kobe" are correlated with semantic distance 1 and "Los," and "New" are removed due to the fact that they belong to stop list. S_{max} is 1.5 here and ambiguous concepts are "Bryant Kobe," "Bryant Mark," and "Reeves Bryant." If we set $\gamma = 0.6$, then the ambiguous concepts "Bryant Mark," and "Reeves Bryant" are discarded since their S_i scores fall below 0.6 ($S_{max} * \gamma = 0.6$). Although, S_i for "New Jersey Nets" is 0.5 which falls below the threshold, we keep it because it is not ambiguous concept.

It might be possible that a relevant concept may be discarded along with irrelevant ones. This is because a given relevant concept may not correlate with other concepts, hence the S_i is low. When relevant concepts are discarded recall will be hurt, because objects with these concepts will not be retrieved if the user request is framed terms of these concepts. For example, the annotated text for an audio object is: "*Flyers* fall to *Leafs*. *Eric* scored two goals and the *Leafs* staved off *Flyers'* third-period rally to hang on for a 4-2 victory Wednesday night over the *Philadelphia Flyers*." The concepts "Desjardins Eric," "Lindros Eric", "Philadelphia Flyers", and "Toronto Maple Leafs" are selected. The S_i propagated scores for these concepts are 1.5, 0.8333, 1.5, and 0.833 respectively. Inter-relationships between player "Desjardins Eric," and team "Philadelphia Flyers" and player "Lindros Eric" and team "Toronto Maple Leafs" are Part-Of. If γ=0.6 is chosen, among two ambiguous concepts "Lindros Eric" will be thrown away, and "Desjardins Eric," will be kept. In other words, the relevant concept, "Lindros Eric" will be discarded.

Note that if there is no correlation, the algorithm fails to resolve ambiguity. In that case, we keep all the selected concepts. For example, the annotated text for an audio object is: "*Young Tiger* hurlers hoping balance offense." Major league baseball's team "Detroit Tigers," and players "Tiger Dmitri," and "Tiger Eric" are selected. The S_i scores for these concepts are 0.5. Due to a lack of correlations, we cannot throw away irrelevant concepts "Tiger Dmitri", and "Tiger Eric." Furthermore, due to the incompleteness of ontology, some irrelevant concepts may be associated with audio objects. For example, the annotated text for an audio object is "Team Up exciting part of NBA All-Star weekend for commissioner. NBA commissioner David Stern believes that Team Up, a program that encourages young people to volunteer their time to the community, is the most exciting part of the All-Star weekend. Former players Bob Lanier and Michael Cooper agree, and say the program is about making a difference in people's lives." Among concepts selected, NBA players "Cage Michael," "Curry Michael," "David Kornel," "Dickerson Michael," "Robinson David," "Wingate David," are wrongly

selected because our ontology does not contain knowledge about NBA commissioner.

One important observation is that when a keyword selects one concept we assume that it is unambiguous, although this unambiguous concept may have a low scores due to no correlation. In the first example of annotated text, as a case in point, one concept, "New Jersey Net" has S_i=0.5. Further, some of these concepts may not be relevant to audio objects. If the annotated text for an audio object is "*Titans* coaches bring game plan to Atlanta. The *Tennesse Titans* fight through the cold of Atlanta and the absence of a bye week to prepare for the SuperBowl against the Rams Sunday. *Titans* quarterback *Steve McNair* believes that the cold weather might actually help his turf toe." Besides, concepts "Mcnair Steve" and "Tennessee Titans," player "Weathers Andre" is selected which is not relevant concept. It may be possible that among ambiguous concepts one of them simply subsumes the other. For example, the annotated text for an audio object is: "*Caps Oates* scores 300th goal; beat Islanders. *Adam Oates* scores his 300th career goal with 5:01 left Monday night, giving the *Washington Capitals* a 3-2 victory over the *New York Islanders*." Concepts "Oates Adam," "Washington Capitals," "York Mike," and "York Rangers" are selected. Note that "York Mike," and "York Rangers" are ambiguous concepts, and the inter-relationship between "York Mike" and "York Rangers" is Part-Of. In that case we discard concept, "York Mike." This is because for this audio object, one team "Washington Capitals" is already selected. Most probably the object conveys information about one team's performance over the other. It is important to note that if concept "York Mike" is selected with higher S_i, we keep this concept.

4.1.1 Further Refinements

Besides regions, several upper level concepts of ontology can be selected. If no region is selected, or a tie between regions occurs for the greatest number of concepts selected, we rely on these upper level concepts. Otherwise, we ignore them for annotation. This is because if a query comes in terms of an upper level concept, that concept will be expanded in terms of more specific concepts by traversing the ontology. In a case in which no region is selected, an audio object will simply be associated with these concepts. In case of a tie, we strive to associate each of these concepts in the selected regions with other upper level selected concepts. Further, if selected upper level concepts hold disjoint properties, these can be used to disambiguate concepts. For example, "basketball," "football," "soccer," and "hockey," are disjoint type concepts. If an upper-level concept is selected which is associated with a concept in a particular region, we keep this region and throw out the others. However, it may happen that several regions may be associated with one of these selected upper level concepts. In that case we cannot resolve the tie.

To illustrate further, if a tie occurs between the regions "NBA", and "NFL", and upper level concept, "Basketball" is selected, and we keep "NBA" and throw out "NFL". This is because the inter-relationship between "NBA" and "Basketball" is Part-Of. On the other hand, "NFL" is associated with "Football" which is disjoint type concept in relation to the concept, "Basketball."

5. Experimental Implementation

In discussing implementation we will first, present our experimental setup, and then present some results of our disambiguation algorithm.

We have constructed an experimental prototype system that is based upon a client server architecture: the server (a SUN Sparc Ultra 2 model with 188 MBytes of main memory) has an Informix Universal Server(IUS) [13], which is an object relational database system. For the sample audio content we use CNN broadcast sports audio [11] and Fox Sports [18]. We have written a hunter program which goes to these web sites and downloads all audio and video clips with closed captions. On average, each of the clips is 1.5 minutes in length. The average size of the closed captions is 112 words. As of today, the total duration of stored audio is 15 hours. Wav and ram are used for media format. Currently, our working ontology has around 7,000 concepts for the sports domain. For fast retrieval, we fetch the upper level concepts of the ontology in main memory while leaf concepts are fetched on a demand basis. Hashing is also used to increase the speed of retrieval. Currently, we have 600 audio clips or objects. Their associated closed captions are used to hook with the ontology.

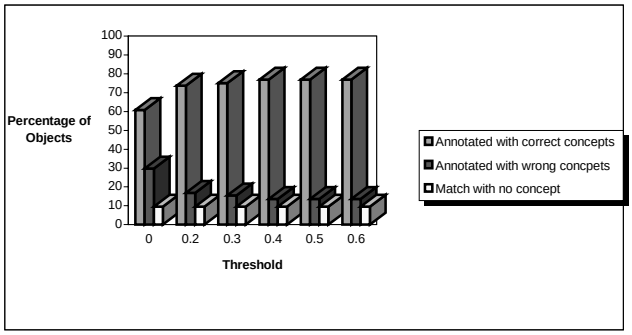
5.1 Results

We have begun study of the performance of disambiguation algorithm by considering what percentage of audio objects it can successfully disambiguate. Furthermore, we would like to study impact of threshold values on pruning irrelevant concepts associated with audio objects and while retaining those which are relevant. We ran our disambiguation algorithms over the audio clips' closed captions. We then inspected the concepts associated with various audio objects. In Fig. 4(a) the X axis represents the value of threshold, γ and the Y axis represents the percentage of instances in which objects are annotated with only correct concepts (category I), with wrong concepts (category II), and with no concept at all (category III). In category II, showing wrong concepts, some correct concepts may also be present (mixed).

For γ =0 when disambiguation algorithm works among different regions, we observed that 9.5% of the objects failed to associate with any concept of the ontology (category III). This is because our ontology is incomplete. For example, an audio object includes reference to a famous hockey player whose career ended ten years ago, and who recently passed away. There is no concept for this player in our ontology, so our algorithm fails to associate a concept with this audio object. Thus, recall will be hurt. On the other hand, 90.5% of the objects are associated with at least some concepts of ontologies (category I & II). Among these, 60.8% objects are all associated with relevant concepts (category I). In other words, in 60.8% of the cases there is no association with an irrelevant concept. Nor in these cases, have we missed any relevant concept. 29.7% objects are associated with at least one irrelevant concept along with relevant concepts (category II). In this case, precision will be hurt due to the annotation of irrelevant concepts. Note that in this case these irrelevant concepts for an audio object are distributed in several regions or a particular region.

With the increasing value of γ, the threshold value, ambiguous concepts will be discarded from category II. Furthermore, this threshold strives to resolve ambiguity for an audio object that exists in a particular region rather than several regions. Recall that an audio object might be associated with several concepts. From there S_{max} score is calculated and ambiguous concepts whose propagated-score, S_i, falls below $S_{max} * \gamma$ are simply discarded. Note that S_{max} varies from object to object. Thus, some objects will be rid of irrelevant concepts and will now be associated with

correct concepts (category I). However, as emphasized earlier, there is a chance that with the increasing value of γ, for a given audio object, we may loose a relevant concept as we shed those which are irrelevant. Thus, recall will be diminished at the expense of improving precision. For a particular γ, in Fig. 4(a) first, second, and third bar represent category I, II and III respectively. Hence, with γ equal to the values 0.2, 0.3, 0.4, 0.5, and 0.6 respectively, 73.7%, 75%, 76.9%, 76.9%, and 76.9% objects are associated with relevant concepts (category I). Further, 16.8%, 15.5%,13.6%, 13.6% and 13.6% of the objects are associated with irrelevant concept(s), along with relevant concept(s), and/or are missing some relevant concepts that are selected a priori (as compared to γ=0). Note also that category III is independent of the increasing value of γ.



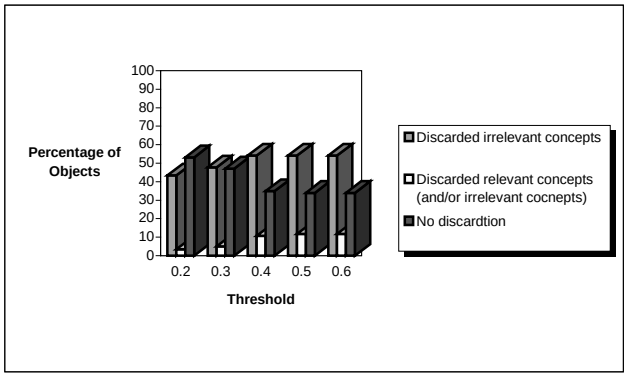


Figure 4 (a) Effect of Threshold on Audio Objects' Associated Concepts (b) Effect of Threshold on Audio Objects' Associated Wrong Concepts

In Fig. 4(b) we show the results of our study of the impact of an increasing value of γ for category II separately. Increasing the value of γ not only leads to the discarding of irrelevant concepts from audio objects but also the loss of relevant concepts. Here, the X axis represents the threshold value of γ, while the Y axis represents the percentage of objects in which discarded irrelevant concepts and relevant concepts occur for category II. For a particular γ, the first, second, and third bars represent the percentage of objects in which all associated irrelevant concept were discarded, the percentage of objects in which at least one relevant concept was discarded, and the percentage of objects in which no ambiguous concept was discarded, out of the 29.7% total objects of category II at γ=0. With γ=0.2, 0.3, 0.4, 0.5, and 0.6, 43.4%, 47.81%, 54.21%, 54.22% and 54.22% of the objects reflect the condition that only irrelevant concepts have been discarded, while only 3.4%, 5.05%, 10.77%, 11.78% and 11.78% of the objects reflect the condition that relevant concepts have been discarded. One important observation is that with an increasing γ, more objects discarded irrelevant concept(s) as compared to a decreasing number of objects in which correct concepts were missed. For γ=0, 60.8% objects are in category I. With γ=0.2, 0.3, 0.4, 0.5, and 0.6, out of 29.7% objects 12.89% (43.4%* 29.7%), 14.20% (47.81%* 29.7%), 16.10%(54.21%*29.7%), 16.10% (54.22%*29.7%) and 16.10% (54.22%* 29.7%) of the objects are all associated with relevant concepts respectively. These will be added to 60.8% objects associated with relevant concepts at γ=0 and are in category I. Thus, with γ=0.2, 0.3, 0.4, 0.5, and 0.6, 73.69%, 75%, 76.9%, 76.9%, and 76.9% objects are in category I respectively in Fig. 4(a).

Note also, with the increasing threshold, γ curves of categories I, and II (in Fig. 4(a)) and all curves in Fig. 4(b) become flat. This is because, at γ=0.4 or higher the disambiguation algorithm is unable to throw any new irrelevant/relevant concepts from category II, since in our data set non-correlated concepts' propagated-scores do not fall into this range. Note that the semantic distance of most of the correlated selected concepts is 1 in our data set. After the propagation of scores among these concepts, their propagated-scores are equal, and they participate in the selection based on maximum scores. On the other hand, the propagated-scores of non-correlated concepts are low.

6. Conclusions

In this paper we have proposed a potentially powerful and novel automatic approach for disambiguation in order to identify appropriate concepts of ontologies from the annotated text of audio. The crux of our innovative algorithm is a procedure for handling concepts in terms of their associations with each other and position within a designated category. Furthermore, for annotated text of audio, this approach strives to discard as many irrelevant concepts as possible while at the same time seeking to maximize the number of relevant concepts retained. We have implemented this algorithm, and observed that up to 76.9% of the audio objects identified in an audio text have been associated with relevant concepts. In the future, we would first like to study the impact of our disambiguation algorithm with regard to user queries in terms of precision and recall. Next, we hope to demonstrate the superior power of ontology over keyword-based search.

7. Acknowledgements

This research has been funded [or funded in part] by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152.

8. References

[1] L. Khan and D. McLeod. Audio Structuring and Personalized Retrieval Using Ontologies. In Proceedings of IEEE Advances in Digital Libraries, Library of Congress, Washington, DC, May 2000.

[2] H. J. Peat and P. Willett. The Limitations of Term Co-occurrence Data for Query expansion in Document Retrieval Systems, J. of ASIS, 42(5): 378-83, 1991.

[3] G. Aslan and D. McLeod. Semantic Heterogeneity Resolution in Federated Database by Metadata Implantation and Stepwise Evolution. The VLDB Journal, the International Journal on Very Large Databases, Vol. 18, No. 2, October 1999.
[4] B. Arons. Speech Skimmer: Interactively Skimming Records Speech. Ph.D. Thesis, MIT Media Lab, 1994.
[5] M. A. Bunge. Treatise on Basic Philosophy:Ontology: The Furniture of the World. Reidel, Boston, 1977.
[6] ESPN CLASSIC. http://www.classicsports.com.
[7] F. Smeaton, and C.J. Van Rijsbergen. The Retrieval Effects of Query Expansion on a Feedback Document Retrieval System, the Computer Journal, 26(3):239-46, 1983.
[8] W. Woods. Conceptual Indexing: A Better Way to Organize Knowledge. Technical Report of Sun Microsystems.
[9] T. R. Gruber. Toward Principles for the design of Ontologies used for Knowledge Sharing. In International Workshop on Formal Ontology, March 1993.
[10] Alexander G. Hauptmann. Speech Recognition in the Informedia Digital Video Library: Uses and Limitations. In 7th IEEE International Conference on Tools with AI, Washington, DC, Nov 1995.
[11]CNN. http://www.cnn.com
[12] Alexander G. Hauptmann and M. Witbrock. Informedia: News-on-Demand Multimedia Information Acquisition and Retrieval. Intelligent Multimedia Information Retrieval, Mark T. Maybury, Ed., AAAI Press, pages. 213-239, 1997.
[13] Informix. Informix Universal Server: Informix guide to SQL: Syntax volume 1 & 2 version 9.1, 1997.
[14] David James. The Application of Classical Information Retrieval Techniques to Spoken Documents. Ph.D. Thesis, University of Cambridge, United Kingdom, 1995.
[15] G. J. F. Jones, J. F. Foote, K. Sparck Jone, and S. J. Young. Video Mail Retrieval. In Proc. ICASSP 95, volume I, pages 309-312, Detroit, May 1995.
[16] A. Hampapur and Ramesh Jain. Video Data Management Systems: Metadata and Architecture in Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media (editors Klas, W. and Sheth, A.), chapter 8, pages 245–285. McGraw Hill, 1999.
[17] D.B. Lenat and R.V. Guha. Building Large Knowledge-Based Systems: Representation and Interface in the CYC Project, Addison Wesley, Reading , MA, 1990.
[18] FoxSports. http://www.foxsports.com
[19] G. Miller. Wordnet: A Lexical Database for English. Communications of CACM, November, 1995.
[20] Large Resources Ontologies (SENSUS) and Lexicons. http://www.isi.edu/naturallanguage/projects/ONTOLOGIES.html
[21] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with WordNet Synsets can Improve Text Retreival, Coling-ACL'98 Workshop: Usage of WordNet in Natural Language Processing Systems, pp 38-44, August 98.
[22] L R Rabiner and R.W. Schafer. Digital Processing of Speech Signals., Prentice Hall, 1978.
[23] G. Salton and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill. 1983.
[24] L.D. Wilcox and M.A. Bush. Training and Search Algorithms for an Interactive Wordspotting System. In Proceedings of ICASSP, volume II, pages 97-100, San Francisco, 1992.
[25] Martin Wechsler and Peter Schuble. Metadata for Content-based Retrieval of Speech Recordings in Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media, chapter 8, pages 223–243. McGraw Hill, 1999.
[26] Ellen Voorhees. Query Expansion Using Lexical-Semantic Relations. 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,1994.
[27] A. F. Smeaton and A. Quigley. Experiments on Using Semantic Distances between Words in Image Caption Retrieval. 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995.