Joint analysis of audio and text data in CASAM

Sergios Petridis, Katerina Papantoniou, Theodoros Giannakopoulos, Georgios Paliouras, Miltiadis Koutsokeras, Elias Zavitsanos, George Tsatsaronis, Giorgos Akrivas, Basilis Gatos, Kostas Ntirogiannis, and Stavros Perantonis

NCSR "Demokritos", Aghia Paraskevi 15310, Greece

Abstract. This paper presents the approach used to extract information from multimedia in the context of the Computer-Aided Semantic Annotation of Multimedia (CASAM) system. In particular, we first describe from a system' s perspective the relevant component of the system, named Knowledge Driven Multimedia Analysis (KDMA) component. We then focus on a particular methodology that allows to improve detection of information found in audio stream of a document, using information found in related text data, provided either as auxiliary sources, speech or user annotations. The methodology is based on separately analysing each medium and then learn a mapping among concepts found in audio and text. This mapping is later used to propose priors for audio classes at the document level and use them to adapt the audio classes posteriors. The evalution results of the described analysis methods on a multimedia news items corpus demonstrate the usefulness of the approach.

Keywords: multimedia analysis, audible events detection, text semantics, fusion

1 Introduction

With the huge increase of multimedia content during the last years, a number of methods have been proposed for automatic characterization, searching and retrieval of this content. Especially for the case of multimedia files from news broadcasts, the usefulness of an automatic content recognition method is obvious: people working in the media, could, for example, more easily search, retrieve and edit news videos, according to some content of interest.

This paper describes the approach proposed and implemented at the KDMA (Knowledge Driven Multimedia Analysis) module of the EU Framework 7 project Computer-Aided Semantic Annotation of MUltimedia (CASAM) system [**?**]. KDMA is the back-end component of the CASAM annotation tool responsible for the low level analysis of multimedia content. It integrates methods that allow extracting information from audiovisual streams and texts, which can ultimately ease the users annotation task. The module works in cooperation with the RMI (Reasoning for Multimedia Interpretation) component and the HCI (Human Computer Interaction) component, which utilize the extracted information and also provides knowledge that facilitate multimedia analysis.

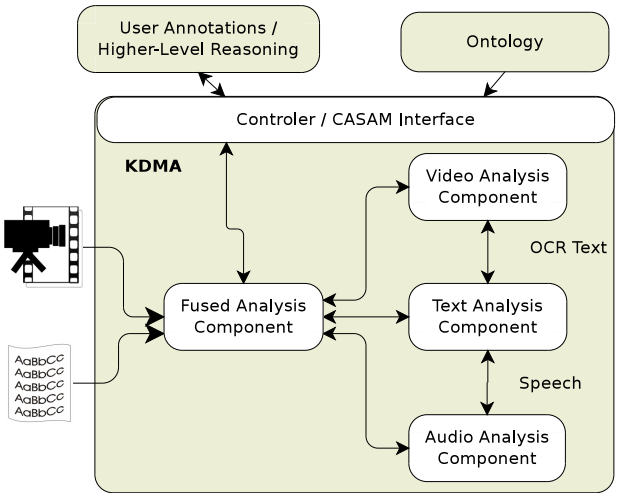

Fig. 1. Internal KDMA Structure.

KDMA includes a large number of methods to deal with particular aspects of multimedia analysis, aiming to provide information in two directions. First, extracting information with respect to people appearing in the video, using speaker clustering, person face grouping and named entity recognition techniques. Second, semantically analyzing the content of the documents, providing information with respect to particular concepts that relate to the analyzed document. KDMA is a complex system designed with an open architecture in mind. Every functionality is implemented as a standalone module and is provided through a common interface. KDMA processes and extracts information is a streaming fashion. Data is processed as soon as it is available and the resulting information is constantly enriched. KDMA employs parallelism to speed up the analysis of documents. This design helps the system to reach almost real time processing of multimodal data and keeps a good balance between computational complexity and speed of execution.

Besides medium-specific methodology, KDMA builds on algorithms that exploit the synergy of media. This paper focus on a particular methodology that allows to improve detection of information found in audio stream of a document, using information found in related text data, provided either as auxiliary sources, speech or user annotations. The methodology is based on separately analyzing each medium and then learn a mapping among concepts found in audio and text. This mapping is later used to propose priors for audio classes at the document level and use them to adapt the audio classes posteriors.

The paper is organized as follows. Section 2 gives an description of KDMA from a system's design perspective view. Section 3 analyses the proposed methodology to detect audio events based on audio and text analysis. In particular, Section 3.1 explains the principle of the approach, Section 3.2 analyses our method-

ology for semantically analyzing textual sources of information, Section 3.3 outlines the base methodology for audio events detection and Section 3.4 explains how the audio classes posterior probabilities are adapted based on text-based analysis audio priors. Finally, Section 4 presents our evaluation results and Section 5 the conclusions.

2 System design

This section aims to describe KDMA from a system design perspective. KDMA is a standalone program, optinally running as web service, capable of processing multiple documents in parallel. KDMA integrates separate analysis components communicating with through a controller. Figure 1 depicts an overall diagram of KDMA architecture. KDMA uses CASAM interface methods and objects to receive input and send results in the form of ontological assertions according to CASAM ontology. The requests for analysis are converted into internal structures and dispatched to media analysis components (Audio, Text, Video and Fusion). The components then produce a series of tags that represent the information detected in the related multimedia documents. Information found is communicated among component and combined through a specific "fusion" component and all results, in the form of ontological assertions are sent to the CASAM system.

2.1 Components and interconnections

The Controler KDMA has a central module, the KDMA Controller, which act as the connecting part between the CASAM architecture and the dedicated analysis modules. In short it implements the following features: handling of Web Service Interface methods. control and dispatching of jobs to multiple threads for parallel processing, the passing of requests to particular KDMA modules and the accumulation of generated results and the periodical sending of them to the CASAM system.

The Components On the other hand, each particular KDMA analysis component (Audio Analysis, Text Analysis, Video Analysis and Fusion) is a separate software library linked into the main KDMA executable program. Nevertheless, the interface between the KDMA Controllerand modules is a commmon abstract class defining the methods and objects all modules should implement and process respectively. This design allows for hiding all common dispatching/scheduling implementation inside to the KDMA Controller. In particular, each component may be asked to process a particular document, take into account particular information with respect to the document and prioritize a particular part of a document to promptly hand out the analysis results, depending on the user's focus.

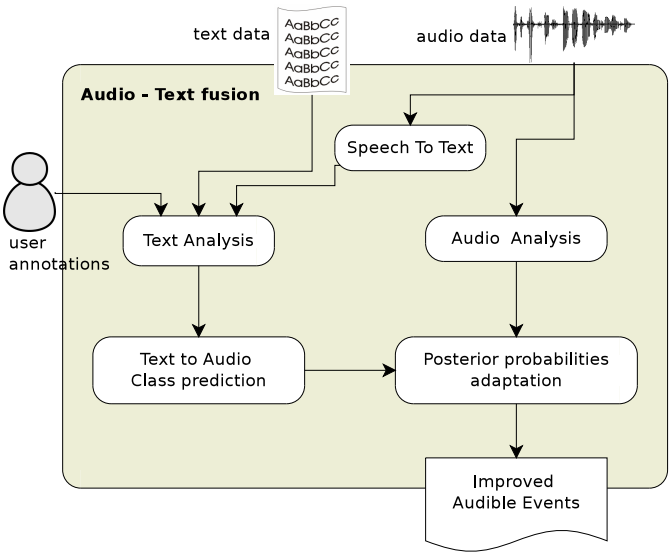

Fig. 2. Schematic diagram of the methodology for improving audio events recognition from speech and text input

Speech and VOCR text An important feature of KDMA, is the analysis of spoken speech and video text through using the text analysis module. Namely, the Audio analysis module locates and extracts speech segments from the audio stream and sends them for speech recognition. The recognized speech segments are then input to the text analysis module. The same pattern is implemented for the VOCR extracted text segments.

Ontological Validation of output Besides KDMA, CASAM integrates other components that are strongly based on description logics formalism. Therefore, results are communicated in form of ontological assertions, represented as RDF-like triples. KDMA also supports the exporting of the extracted information in OWL format. The exported OWL documents are then validated with reasoners like Pellet [**?**] and RacerPro[**?**]. The validation is done off line and is mainly using for monitoring/debugging the system, such that the KDMA results are guaranteed to be compliant with the ontology used in CASAM.

2.2 Features

KDMA is supporting an incremental communication mode for both receiving request and sending the results back. This allows to handle:

Streaming information Answer to query may given back in blocks. E.g. if a 5 minute video is given for analysis, the first reply of KDMA may concern the first minute, the second one the second minute, etc.

Level of Granularity For the same query, easy to extract information by KDMA is sent first, and thus fast, while hard to extract information is sent later.

Adapting to user feedback When knowledge regarding the analyzed document is changed, either by user-provided feedback or by higher level reasoning, KDMA re-analyses the data and provide updated analysis results.

3 Joint analysis of multimedia data

Working simultaneously with multiple media is an advantage, when done properly. In several case, the same information is repeated in several forms, in more than one media. Thus, extracting semantics can be done with greater confidence and accuracy once all media are considered. In this section, we describe a methodology which is part of the KDMA module aiming at improving audiovisual events detection, through analysis of vocr. speech and/or available related text documents. The description wil focus on audio-text information fusion, though the extention to video - text information fusion is straightforward. Generaly, this methodology may be applied in cases, where the expected coupling between different media is loose. The methodology allows also to graciously take account information provided by the user.

In what follows, we first given an overview of the fusion methodology (Section 3.1). We then describe modality-dependent algorithms that extract information from text (Section 3.2) and audio (Section 3.3). Finaly, in Section 3.4 we show how extracted information from text data is used to improve the accuracy of information extracted form audio data by adapting the posterior probabilities.

3.1 Fusion methodology

High level Fusion Fusion approaches that may be used to jointly extract information from multimedia data are generally categorized in three types: low-level fusion, mid-level fusion and high-level fusion. These differ at the level of representation in which fusion takes place.

In low-level fusion, sources are required to be *homogeneous* i.e. of the same type, so that one may take advantage of very specific techniques, developed for the specific data structure and data structural elements. This enables the use of medium-specific algorithms, to be applied on the aligned data. In mid-level fusion, the considered media, although not homogeneous, they are somehow alignable, i.e. one can establish a correspondence between their structural elements. A typical case of alignable data are audiovisual documents, where video and audio are aligned on their time dimension. A possible approach here is to homogenize the elements by extracting features from all different media. Since feature-vectors are medium-neutral, they can be concatenated, thus producing a unified feature vector for the joint multimedia element.

Still, there are cases, in which multimedia data are neither homogeneous nor alignable. Then, a way to proceed is to defer the fusion between corresponding elements to a common, medium neutral, symbolic level. This implies that

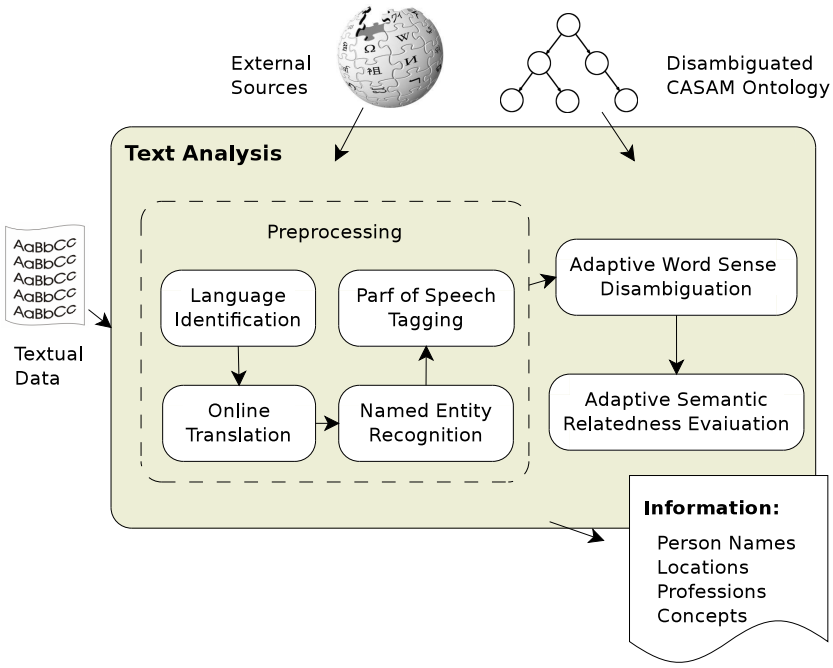

Fig. 3. Text Analysis Methodology.

a complete analysis is done in each medium, aiming to achieve information at a symbolic level of representation prior to fusion. This method is referred to as *high-level* fusion, though the terms *symbolic-level* fusion is sometimes used. A characteristic of this approach is that it does not impose extracting feature vectors from each medium. but allows for particular methods to be applied independently until the symbolic level is reached.

An important advantage of high-level fusion is that it can handle the case of weaker alignment of documents. In particular, it is possible that alignment of documents can be done in coarse documents elements, within which the generated features by each medium can change significantly. The methodology we present here falls into the high-level fusion type, since it relies on symbolic results obtained by separately analyzing each medium at the document level.

A further advantage of high-level fusion is that it can accounts graciously for information directly provided by the user, as in the case of CASAM. Indeed, it is much more likely that the user provides information by means of high-level concepts, rather than with numerical feature values. The method presented here allows to directly take into account user-provided information in terms of priors over concepts.

Predicting Audio Priors based on Text posteriors An overview of the methodology is summarized in Figure 2. In particular, an audiovisual document

along with textual information is separately analyzed through the the text and audio analysis modules. Note that, within the textual input, extracted speech from the audio channel is also taken into account. At the same time, a user may provide additional information with respect to which concepts are important in the document, in terms of concept priors.

By considering the (separate) analysis results of audio and text data at the document level, one may effectively create a mapping from text-extracted to audio-extracted concepts. In other words, one may try to predict the probabilities that specific audible events occur in the document, taking into account concepts existing in textual data referring to the same document, without taking into account audio analysis. These priors may then be used to improve the final prediction of the audible events detection algorithm.

In practice, by assembling all these mappings for the documents of a given corpus, a reference set is created. This set is used to learn a regression model that predicts, given as input the priors for text, the priors for audio concepts for a particular document. These priors are then used to adjust the posteriors for audio classes, given particular audio segments, as described in Section 3.4.

3.2 Semantic analysis of textual data

Textual data are important sources of information, particularly for recognizing topics of discussions, and particular named entities, such as names, locations and organizations. Within the CASAM system, text data may be provided in several forms, such as when given directly by the user, or as the result of speech transcription of the audio data. It may also be found in written form on the video frames. In this case, an important step is to locate the text on the video frames. Once textual data are available in direct form, an elaborate methodology for their semantic analysis takes places, as detailed in Section 3.2.

The Text Analysis component comprises several sub-components to as depicted in Figure 3. Some components perform pre-processing tasks while other perform semantic annotation of text with ontology concepts using information from the pre-processing phase.

Preprocessing Steps When the analysis of text is requested, the first step is the identification of the text language. To that end, an N-gram-based approach for text categorization is used [**?**]. The approach is based on identifying N-grams whose occurrence in a document gives strong evidence for the classification of a text under a particular category(language). If, the language is other than English, then a text translation is performed in order to align the input text to the language that is used by the rest of the components and the ontology. For the automatic translation of the input text to English, the Google Translate Service is used.

Subsequently, a named entity recognition component allows the system to identify entities in the input text and semantically annotate them with the corresponding concepts of the given ontology. The OpenCalais service ([**?**]) has been

selected for this purpose. OpenCalais is a popular, free extraction service. It can identify a wide range of entities as well as commonplace relations between those entities. Finally, we apply a part of of Speech Tagger, in order to assign parts of speech tags to each word such as verb, noun, adjective etc. For this task the maximum entropy-based part of speech tagger of Stanford ([**?**]) is used. The information of the POS tagger is mainly used for word sense disambiguation.

Preprocessing also includes word sense disambiguation, i.e. the task of resolving semantic ambiguity pervasive in natural language. The aim of word sense disambiguation methods is to identify the most appropriate meaning for any given word with respect to its context. In CASAM the output of the word sense disambiguation process is exploited in the phase of semantic relatedness calculation. The calculation is performed between the specific meaning of a word and an ontology concept, providing a more accurate score. The disambiguation is performed through a state-of-the-art unsupervised method which expoits WordNet([**?**]). In parallel, with the plain word sense disambiguation method a biased to the domain WSD method has also been tested. In this method the word sense disambiguation process in enhanced with information about the domain.

Semantic Relatedness Calculation The main component in the sequence performs the semantic annotation of the input text with ontology concepts. The main idea is the calculation of a degree of semantic closeness between text keywords or keyphrases and ontology concepts. This degree takes a value in the interval [0, 1] with high values indicating close semantic relation. The annotation procedure comprises three consecutive steps: exact matching, stem matching and semantic matching. In the last step more sophisticated techniques are employed for the calculation of the semantic relatedness between the terms of the text and the ontology concepts that have not been used in the first more trivial annotation steps. The main approaches developed are presented in brief below.

Baseline This method consults the WordNet, in order to retrieve a list of synonyms for the lexicalization of each concept of the domain ontology. The calculation of relatedness in this method depends to a large extent on the set of the retrieved synonyms. In particular, it tends to assign high relatedness scores in cases where the semantic distance between a concept and its synonym is small, and lower relatedness scores otherwise. Details and evaluation results for this method are decribed in [**?**].

Omiotis The Omiotis ([**?**]) method also uses WordNet to perform the semantic annotation by applying the Omiotis measure for the calculation of relatedness between two terms i.e. a candidate word and the lexicalization of a concept. The superiority of this measure lies in the utilization of all the provided semantic relations by WordNet, and that it can be applied to terms of any POS type. The measure provides the highest correlation with human judgments among the dictionary-based measures of semantic relatedness.

Relatedness-based Annotation with Omiotis and Wikipedia This method employs an additional measure based on Wikipedia, in order to handle those cases not supported by Omiotis i.e. the pairs of words that do not appear in WordNet. The method tries to augment the semantic annotations based on the assumption that if something does not exist in WordNet, it may be found in Wikipedia (for example proper nouns). The measure of Milne and Witten ([**?**]) is employed for this purpose which is the fastest among several alternatives and provides high correlation with human judgments.

Methods based on collaborative resources In these methods, we investigate whether and how the use of different collaborative lexical sources (apart from Wikipedia) can prove beneficial for the semantic annotation task. For this purpose, we extend our so-called baseline method, which consults the WordNet, so as to be applied to other two community-based lexical resources which in recent years gain considerable attention from researchers. The resources are the Wiktionary ([**?**]) and DBPedia ([**?**]). Wiktionary is a collaborative project to produce a free-content multilingual dictionary while DBPedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. The structure of Wiktionary bears significant resemblance to that of WordNet while DBPedia is organized in the form of RDF datasets.

As with the baseline method the semantic relatedness between the lexicalizations of ontology concepts with terms of the input text is calculated based on the size of the retrieved synonyms and related terms. In the case of Wiktionary we retrieve the synonyms, the related terms and the derivatives of a word while in the case of DBPedia we retrieve information for the **"redirect"** and the **"category"** dataset.

3.3 Base methodology for audible events detection

The base method for detecting audible events is summarized in Figure 4. In this section, we present the main characteristic of the approach. For a more detailed description, the reader is referred to [**?**].

Targeted audible event classes To begin with, an audio class detected by the audio module is *speech*. Speech tracking is useful for both detecting different speakers [**?**] and extract the speech transcript, which is then forwarded for semantic analysis. Other than that, the following more semantically rich audible event classes have been considered in our study, mainly because of their adequate representation in the studied CASAM corpus: music,water,wind,engine,applause

Given the nature of the signals under study, the audio events, most of the times, exist as background events, with speech being the major sound. Therefore an audio segment can, at the same time, be tagged as speech and as some other type of event (e.g. wind).

Short-term feature extraction For detecting the audio events, a 21-D feature vector is extracted for each audio segment. Towards this end, the audio stream

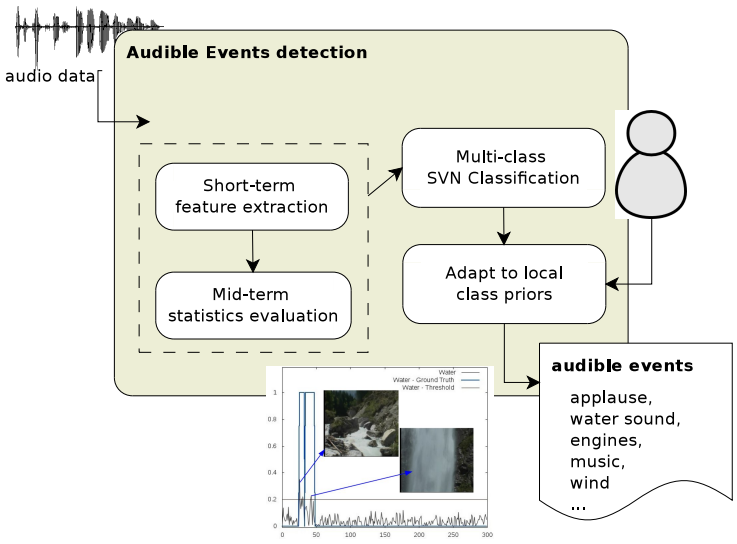

Fig. 4. Recognizing audible events.

is first short-term processed, i.e., the signal is split into short-term windows (40 msecs long with 50% overlap) and for each window seven audio features are extracted. In particular, the following features are calculated for each short-term window: Energy, Zero Crossing Rate, Energy Entropy, Spectral Centroid, Position of the Maximum FFT Coefficient, Spectral Rolloff, Spectral Entropy and Audio event detection

mid-term statistics evaluation The above process leads to a representation of the audio stream based on seven feature sequences. In the sequel, a mid-term analysis approach is adopted, according to which, three statistics are computed, on a mid-term basis, for each of the seven feature sequences. The three statistics are: mean value, standard deviation and std by mean ratio. The selected mid-term window size is 2 secs, with an overlap of 50%. According to this, each 1-sec segment of the audio stream is now represented using a 21-D feature vector. This vector is used in the next stage, in order to detect audio events.

Event detection In order to classify each 1-sec audio segment (represented by a 21-D feature vector, as described above), we have adopted Support Vector Machines (SVMs) and a variation of the One Vs All classification architecture. Towards this end, each binary classification subtask ,e.g., 'Speech Vs Non-Speech', is modeled using a separate probabilistic soft-output SVM. For each audio segment, as long as the six SVM outputs are computed (one for each audio class), a thresholding criterion is applied on each SVM soft output. In the sequel, the soft-outputs that correspond to non-speech events and which have survived from the thresholding process are compared and the class with the maximum respec-

tive soft output is kept as a result for the current audio segment. For the case of speech, only the thresholding criterion is applied on the corresponding SVM output.

3.4 Adapting audio posteriors using text-based audio priors

Till now, we have discussed how to obtain probability for classes related somehow to a multimedia document by text and audio analysis, as well as how to predict the probabilities of audio classes, without analysis the audio, given the posterior probabilities of classes extracted by text analysis.

In a way, predictions of audio classes from text classes, are "better" prior probabilities for audio classes. In deed, they can be considered as priors, since they are provided without considering audio data. Also they are "better" priors, in the sense that text are more related to a particular document than priors obtained while training the audio classifier using the complete corpus. Therefore, it make sense to use these new priors in order to improve the final detection accuracy of the audio classes.

Note that, in audio analysis, audio events are extracted by means of a discriminative model, in a particular SVM, as opposed to a generative one. It follows that, an issue that has to be addressed is how to change the posteriors of audio classes, given that the priors have changed, so that the classifier takes advantage of the extra information from text analysis.

Generative models To see how posteriors are adapted in the discriminative case, let us first assume that a generative model is used, which estimates the conditional probabilities of class c given the sample $\mathbf{x}$, based on the Bayes rule. Assuming that the sample and the class take value from the domains $\mathcal{X}$ and $\mathcal{C}$ respectively, this is expressed as follows:

$$\begin{aligned} p_{\mathcal{X},\mathcal{C}}\left(c|\mathbf{x}\right) &= p_{\mathcal{X},\mathcal{C}}\left(\mathbf{x}|c\right) \cdot p_{\mathcal{C}}(c) \cdot \frac{1}{p_{\mathcal{X}}(\mathbf{x})} \\ &= p_{\mathcal{X},\mathcal{C}}\left(\mathbf{x}|c\right) \cdot p_{\mathcal{C}}(c) \cdot \frac{1}{\sum_{c\in\mathcal{C}} p_{\mathcal{X},\mathcal{C}}\left(\mathbf{x}|c\right) \cdot p_{\mathcal{C}}(c)} \end{aligned} . \tag{1}$$

where the probability of generating a specific $\mathbf{x}$ given that the class is c, i.e. $p_{\mathcal{X},\mathcal{C}}\left(\mathbf{x}|c\right)$, as well as the priors over the class $p_{\mathcal{C}}(c)$ are estimated from trained generative models using some reference set.

Assume now that the test sample is drawn from a subdomain of $(\mathcal{X},\mathcal{C})$, where different information about the priors of the class probabilities is available. Let this subdomain be $(\mathcal{X}',\mathcal{C}')$ and let $p_{\mathcal{C}'}(c)$ be the priors for the classes for the subdomain. Given that the class-conditional observation distributions do not

change, the conditional probabilities of classes c may now be evaluated as

$$\begin{aligned} p_{\mathcal{X}',\mathcal{C}'}(c|\mathbf{x}) &= p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c) \cdot p_{\mathcal{C}'}(c) \cdot \frac{1}{p_{\mathcal{X}'}(\mathbf{x})} \\ &= p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c) \cdot p_{\mathcal{C}'}(c) \cdot \frac{1}{\sum_{c\in\mathcal{C}} p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c) \cdot p_{\mathcal{C}'}(c)} \\ &= \frac{p_{\mathcal{C}'}(c)}{\sum_{c\in\mathcal{C}} p_{\mathcal{C}'}(c) \cdot p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c)} \cdot p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c) \end{aligned} \tag{2}$$

Discriminative models In our case, however, generative probabilities are not available, while discriminative are. Nevertheless, the assumption that the generative probabilities stay the same still hold. By combining eq.(1) and (2) we obtain

$$p_{\mathcal{X}',\mathcal{C}'}(c|\mathbf{x}) = p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x}) \cdot \frac{p_{\mathcal{C}'}(c)}{p_{\mathcal{C}}(c)} \cdot \frac{p_{\mathcal{X}}(\mathbf{x})}{p_{\mathcal{X}'}(\mathbf{x})} \tag{3}$$

where $p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x})$, $p_{\mathcal{C}'}(c)$ are known. In particular, let us define, for each class, the *prior probability ratio* as

$$r(c) = \frac{p_{\mathcal{C}'}(c)}{p_{\mathcal{C}}(c)} \tag{4}$$

by means of which, the formula rewrites as

$$p_{\mathcal{X}',\mathcal{C}'}(c|\mathbf{x}) = p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x}) \cdot r(c) \cdot \frac{p_{\mathcal{X}}(\mathbf{x})}{p_{\mathcal{X}'}(\mathbf{x})} \tag{5}$$

Now, $p_{\mathcal{X}}(\mathbf{x})$ and $p_{\mathcal{X}'}(\mathbf{x})$ are not assumed to be known. Nevertheless. we just need to evaluate their ratio $\frac{p_{\mathcal{X}}(\mathbf{x})}{p_{\mathcal{X}'}(\mathbf{x})}$. This can be done as follows:

$$\begin{aligned} \frac{p_{\mathcal{X}'}(\mathbf{x})}{p_{\mathcal{X}}(\mathbf{x})} &= \frac{\sum_{c\in\mathcal{C}} p_{\mathcal{X}',\mathcal{C}'}(\mathbf{x}|c) \cdot p_{\mathcal{C}'}(c)}{p_{\mathcal{X}}(\mathbf{x})} \\ &= \frac{\sum_{c\in\mathcal{C}} p_{\mathcal{X},\mathcal{C}}(\mathbf{x}|c) \cdot p_{\mathcal{C}'}(c)}{p_{\mathcal{X}}(\mathbf{x})} \\ &= \frac{\sum_{c\in\mathcal{C}} p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x}) \cdot p_{\mathcal{X}}(\mathbf{x}) \cdot \frac{p_{\mathcal{C}'}(c)}{p_{\mathcal{C}}(c)}}{p_{\mathcal{X}}(\mathbf{x})} \\ &= \sum_{c\in\mathcal{C}} p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x}) \cdot r(c) \end{aligned} \tag{6}$$

Notice that the ratio of the observation probabilities, before and after the priors knowledge, equals the expected ratio of the class probabilities before and after the priors knowledge, weighted by the probability of each class given the observation. Altogether, the class probabilities given the new priors can be evaluated as:

$$p_{\mathcal{X}',\mathcal{C}'}(c|\mathbf{x}) = \frac{r(c)}{\sum_{c\in\mathcal{C}} r(c) \cdot p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x})} \cdot p_{\mathcal{X},\mathcal{C}}(c|\mathbf{x}) \tag{7}$$

	Baseline	Omiotis	OW	OWSD	OB	Wiktionary	Category	Redirect
DW								
Precision	0.77	0.77	0.77	0.80	0.80	0.83	0.76	0.73
Recall	0.69	0.68	0.69	0.69	0.69	0.78	0.77	0.77
F-Measure	0.72	0.71	0.73	0.73	0.74	0.80	0.76	0.75
LUSA								
Precision	0.37	0.51	0.51	0.54	0.58	0.37	0.56	0.61
Recall	0.70	0.57	0.58	0.55	0.59	0.67	0.63	0.62
F-Measure	0.47	0.51	0.50	0.51	0.58	0.46	0.61	0.63

Table 1. Text analysis Evaluation results for the Deutche Welle and LUSA dataset. OW stands for Omiotis-Wiki, OWSD stands for Omiotis-WSD and OB stands for Omiotis-Biased WSD

which reveals that $\frac{p_{\mathcal{X}'}(\mathbf{x})}{p_{\mathcal{X}}(\mathbf{x})}$ can also be considered as a normalizing term, guaranteeing that

$$\sum_{c \in \mathcal{C}} p_{\mathcal{X}',\mathcal{C}'}\left(c|\mathbf{x}\right) = 1$$

4 Evaluation

Text Analysis We now presents the empirical evaluation of our semantic annotation methods in two datasets. The first dataset comprises 51 documents provided by the LUSA Agency and the second comprises 64 documents provided by the Deutche Welle. Both datasets were manually annotated with the corresponding ontology concepts to a gold standard. Performance measured in terms of Precision, Recall and F-measure, which are standard measures from the field of Information Retrieval. Table 1 present the results of the various approaches.

A first conclusion from the experimental results is that the baseline method along with the methods based on the collaborative resources outperform. This behavior was expected to some extent, since the rest of the methods rely on the calculation of the relatedness in order to perform the annotation. In particular, in case where the relatedness value between a candidate keyword and an ontology concept is greater than zero, then this keyword will be annotated, even if the relatedness value is very close to zero, which would normally signify that it should not be annotated. A second conclusion is that the use of different lexical sources can lead to an increase in the F-measure. However, more experiments must be performed in different domains apart from the environmental and in larger datasets, for a broader assessment of the performance. In addition, the most interesting issue which is our immediate future work is the effective combination of these measures.

Class names	Recall(%)	Precision(%)
Speech	82	90
SoundofAir	20	82
CarEngine	42	87
Water	52	90
Music	56	85
Applause	59	99
Average (non-speech events)	45	86

Table 2. Performance measures for the base audio event detection module.

Based Audio Analysis The base audio event detection method has been evaluated using two datasets, which have been populated in the CASAM project: one from the German international broadcaster (DW - Deutsche Welle) and the second from the Portugese broadcaster (Lusa - Agncia de Notcias de Portuga). Around 100 multimedia streams have been manually annotated, with more than 7 hours of total duration.

The recall and precision rates for the event detection task have been defined as performance measures for the given task (see [**?**] for more details). In Table 2, the results of the event detection process are presented. It has to be noted, that the thresholds used in the event detection stage, have been estimated so as to lead to high precision rates. This was obviously achieved, as all of the audio events were detected with a precision rate higher than 80%.

Fused media analysis To evaluate the proposed methodology for improving audio events recognition based on speech and/or text information, we have measured the quality of estimating audio classes priors for particular videos, given accompanying auxiliary text and/or speech extracted from the same video.

Namely, 75 audiovisual news items from DW and LUSA together with auxiliary text material have been used as input. Using the text analysis module,, a text analysis for each item has been conducted, resulting in 146 distinct instantiated concepts (such as . Agency, Agriculture, Air, Animal, Applause, Association, Audience, Bank, Banner, Bike, Biofuel, Biomass, Bird ...). On the other hand, 9 different audio concepts have been instantiated, the 9th one being a general concept "other". which acts as a general container for concepts with very few samples.

Using these data, 9 different sample sets, one for each audio class, having 75 146-dimensional samples each, have been created. A 10-fold cross validation test has been conducted to a number of classifiers, for each one of these sample sets. The results are shown in Table 3. These are SMO [**?**], Gaussian Process [**?**], K-Star [**?**] and a decision stump, which does regression based on mean-squared error.

By looking at the results, one may say that text information, provided as features to the classifier, allows to predict significantly the a-prior probability

Audio Class	Baseline	SMO	Gauss.	K-Star	Dec. Stump
Applause	100.00	59.24 •	93.43 •	58.68 •	75.59 •
Engines	100.00	89.85 •	99.41	124.06 ○	106.37 ○
Music	100.00	92.05 •	108.41 ○	171.15 ○	96.20
Twittery	100.00	78.14 •	93.34 •	92.93 •	82.97 •
Water	100.00	65.28 •	90.99 •	63.85 •	137.25
Wind	100.00	78.61 •	98.54 •	102.21	122.13
Noise	100.00	93.50 •	98.87 •	98.78	110.99 ○
Speech	100.00	97.86 •	103.22 ○	135.15 ○	95.57 •
Other	100.00	91.50 •	100.02	113.14 ○	101.68
Average	100.00	82.89	98.47	106.66	103.20

○, • statistically significant improvement or degradation

Table 3. Results of Audio classes Prior improvements. Numbers correspond to relative error in prior estimation with respect to the baseline approach, which is using the corpus-wide prior probability.

of an audio class in the related audio material. Although a large variability of results exist between classifiers, the SMO classifier manages to significantly decrease the error estimation, with respect to general priors of classes, for all audio classes. Note in particular the Applause and Water classes, which prior estimation is improved by ∼40 percent. These results are very encouraging with respect to predicting audio concepts from detected text concepts.

5 Conclusion

This paper has presented the KDMA component of the CASAM prototype system from a system's design perspective and focused on a particular fusion methodology of the component suitable to combine information from audio and text, in order to improve the accuracy of the audio events detection.

The proposed methodology enabled using dedicated algorithms for extracting information from each medium separately and independently. Improved methods for audio or text analysis may thus be used in place, as pluggable components. The methodology handles especially the case of soft discriminative classifiers, where the probabilistic output is adapted by more accurate class priors.

The core of the methodology consist of a regression model that maps class probabilities of one medium to class probabilities of the other. Evaluation of generalization accuracy of several possible model, including SVM, show significantly positive results in providing better estimates for the audio class priors.

Overall, a large number of methodologies are put in place in KDMA, including speaker clustering, face detection, spatially locating text in videos and many others, which are still under development during the last phase of the CASAM project. The on-going evaluation of the KDMA component within the CASAM

system by end users with respect to both leveraging manual annotation effort and enrich the annotation results.