Ontology-Based Retrieval of Human Speech

Javier Tejedor[1], Roberto García[2], Miriam Fernández[1], Fernando J. López-Colino[1], Ferrán Perdrix[2,3], José A. Macías[1], Rosa M. Gil[2], Marta Oliva[2], Diego Moya[1], José Colás[1], and Pablo Castells[1]

[1]Universidad Autónoma de Madrid
C/ Tomás y Valiente 11
28049 Madrid, Spain
{javier.tejedor,miriam.fernandez,
fj.lopez,j.macias,diego.moya,
jose.colas,pablo.castells}@uam.es

[2]Universitat de Lleida
C/ Jaume II 69
25001 Lleida, Spain
{rgarcia,rgil,
oliva}@diei.udl.es

[3]Diari SEGRE
C/ Del Riu 6
25007 Lleida, Spain
fperdrix@diarisegre.com

Abstract

As part of the general growth and diversification of media in different modalities, the presence of information in the form of human speech in the world-wide body of digital content is becoming increasingly significant, in terms of both volume and value. We present a semantic-based search model for human speech corpora, stressing the search for meanings rather than words. Our framework embraces the complete recognition/retrieval cycle, from word spotting to semantic annotation, query processing, and search result presentation.

1. Introduction

Along with the general growth and diversification of media in different modalities (text, image, graphics, video, audio), the presence of information in the form of human speech in the world-wide body of digital content is becoming increasingly significant, in terms of both volume and value. For instance, in the news domain, news companies are turning into news media houses, owning radio stations and video production companies that produce content not supported by the print medium, but which can be delivered through Internet newspapers. Broadband networks and streaming technologies enable efficient access to audio and video contents for everyone on the WWW. Most radio stations are offering live broadcasting services, and online audio archives on the net. Even at the individual level, the amount of personal recordings that a single person can accumulate, and share with his kin and kith over the years is becoming quite considerable, to the point of raising important management difficulties.

Such new perspectives in the area of digital content call for a revision of mainstream search and retrieval technologies, currently oriented to text and based on keywords. Classic Information Retrieval (IR) has been oriented to text content for several decades. The turn of the current decade has brought a raising interest (both as a research problem and a business opportunity), and investment, in image, audio, and video search technologies, which can be said to be still incipient but making fast progress. However, the same as traditional IR is based on character strings (keywords), current multimedia retrieval approaches mainly rely on low-level content features (pixels, media descriptors), or text (e.g. image caption, audio transcription).

The research presented here aims to the enhancement of current IR technologies in two directions, namely, a) the development of semantic-based retrieval technologies that support search by meanings rather than keywords, providing users with more powerful retrieval capabilities to find their way through in increasingly massive search spaces; and b) the integration of human speech processing technologies in the semantic-based approach, extending the semantic retrieval capabilities to audio content. The research is being undertaken in the frame of a national project[1] involving two universities and a Spanish news media group.

2. State of the art

Search engines for images and music have been around for several years now (e.g. Google image search, AllTheWeb, Kazaa, AltaVista, Lycos). Video (e.g. Google video search) and voice (e.g. Speechbot, Altavista, Singingfish) are also being addressed, but on a more incipient lane. Most radio stations offer their audio files on the Web, and resources in audio format proliferate in libraries and other collections.

The same as text retrieval techniques are based on the analysis of word occurrences in free text documents, search in spoken corpora rely on the capability to detect

[1] See http://nets.ii.uam.es/s5t

words in a continuous speech stream. The main challenge is to achieve the highest keyword detection rate while minimizing the false acceptance rate. Most of the developed wordspotters use variants of Hidden Markov Models (HMM) for continuous speech recognition [6], [8]. In such systems, the non-keyword intervals are represented by a variety of filler models, varying from a few phonetic or syllabic fillers to whole words [2]. The use of language models for the transitions between keywords and filler models has also been explored [6].

Those systems share with mainstream text IR systems a common limitation, namely, their ability to represent meanings is based on counting word occurrences, regardless of the relation between words [12]. Most research beyond this limitation has remained in the scope of linguistic [13] or statistic [3] information. On the other end, IR is addressed in the Semantic Web field from a much more formal, we might say ideal, perspective. In the Semantic Web vision, the search space consists of a totally formalized corpus, where all the information units are unambiguously typed, interrelated, and described by logic axioms in domain ontologies, in such a way that a query has either a perfect, exact answer, or none at all [9].

In our view, it is not realistic to expect that the massive content flows and spaces where people develop their activity on a daily basis (such as the Web, large intranets, or even the personal desktop) can be fully represented in that way. Thus, we propose a hybrid approach, where ontologies and unstructured contents coexist, in such a way that available domain ontologies and knowledge bases (KBs) are exploited to provide better answers to user queries, but the results consist of content fragments (text or speech) rather than ontological data, selected and ranked by gradual (rather than boolean) predicted relevance, in the common IR vs. data retrieval view [12].

3. Ontology-based retrieval

Our approach assumes the availability of domain KBs where concepts of several domains (art, economy, politics, sports, etc.) are formally described by means of ontologies. These concepts may appear in the content to be searched for. For example, for the archive of a sports newspaper, the KB would contain information about sportsmen, clubs, competitions, etc., including data related with sports events, records, or personal information, as well as relationships like who coaches a team, or who are the participants of a competition.

Our second assumption is that the content corpus to be searched is annotated by KB elements. This provides the key to bridge the ontology-based query technologies to the IR indexing and ranking strategies, as we shall explain later. For example, the instance describing a sports club in the KB can be associated to the news where the club is mentioned. The problem of the (manual or automatic) semantic annotation of text documents has been largely addressed in the Semantic Web field [11]. Our research includes a proposal to produce content annotations in a semi-automated way, addressing the additional problem that the discourse to be annotated is provided in audio format.

Our third and last assumption is the availability of a concept-keyword mapping where each KB item is associated to one or more string keywords (or phrases), which represent the textual form under which the concept commonly appears in a free text or speech. Obtaining this mapping is not a trivial problem, but it is not the focus of this paper. Instead, the results of prior work that addresses this problem [11] have been reused in our experiments.

3.1 Semantic annotation

Our approach to semi-automatic content annotation is based on the concept-keyword mapping described above. Specifically, when some of the textual forms of a concept is uttered in a piece of speech, the content is annotated with the concept. Polysemic words and other ambiguities are treated by a set of heuristics, the description of which can be found by the reader in [1]. Two problems remain to be addressed: how to find keywords in speech, and how to assess the importance of different concepts occurring in a speech fragment. The former is addressed in Section 4, and the latter is solved as follows.

In classic IR models, keywords appearing in a document are assigned weights which account for the fact that some words are better representatives of documents than others. Similarly, in our model, annotations are weighted according to the importance of the concept for the document meaning. Weights are computed automatically by an adaptation of the TF-IDF algorithm [12], based on the frequency of concepts in documents, where the “occurrence” of a concept is primarily defined as the utterance of some of its textual forms.

3.2 Semantic query processing

Our approach for the execution of queries can be seen as an evolution of the vector-space IR model, where the keyword-based index is replaced by a semantic knowledge base. Our system takes as input a formal ontology-based query, for which SPARQL and RDQL are currently supported. The problem of providing the user with friendly query interfaces is not a trivial one, and is addressed in Section 6 of this paper. The formal query is executed against the KB using a state-of-the art query engine, which returns a list of instance tuples that satisfy the query. Unlike the common approach in the Semantic Web vision [9], this is not the final result. Instead, the last step consists of returning the contents that are annotated with the values from the formal results.

The contents are ranked by a standard IR algorithm, based on a cosine similarity function to compare the query vector and the document vector. The basis vector space in which queries and documents are represented is defined by the ontology concepts and KB instances, in a way that each concept and each instance define an axis of the vector-space. The document vector is then defined by the weights of the conceptual annotations. In principle, the query vector is Boolean, where the instances that appear in the formal result set have a weight of 1, and the rest have a weight of 0. In practice this is refined in a way that instances that appear in multiple result-set tuples are given a higher weight.

4. Speech recognition

The keyword spotting system is the piece that completes the approach described so far. It provides the capability to detect and count concept occurrences by finding their associated keywords in the audio corpus. The keyword spotting system takes as input the list of all keywords from the concept-keyword mapping, and returns an XML file containing the set of keywords recognized and the time intervals where these keywords were located. The details of these techniques are described next.

The speech recognition system is first trained with an annotated corpus, after which it is ready to be used on the target speech corpus. The speech is parameterized using Mel-frequency cepstral coefficients (MFCC). In addition to the 13 static cepstral coefficients (including the 0th coefficient), deltas and accelerations are computed to generate 39-dimension observation vectors. Four different acoustic models, to be used during the speech recognition processes, were defined, based on the Spanish Albayzin database [10]:

- Allophone Models (AM): 47 models were trained taking into account the different phonological rules in the Spanish language.
- Phonemes Models (PM), including only the 23 phonemes in the Spanish language.
- Broad classes Model (BM), grouping the 23 phonemes into eight classes: nasals, closed vowels, opened vowels, median closed vowels, deaf fricatives, deaf explosives, sound explosives, and liquids.
- Average Phonemes Model (APM): a unique background model was trained considering a single phoneme in Spanish.
- An initial (at the beginning of each sentence), final (at the end of each sentence), and short silence (between words) were added to each configuration.

As is shown in Figure 1, the keyword spotting tool is based on a hybrid word/phoneme architecture where two different recognition processes take place: phonetic decoding and keyword spotting. The former recognizes a set of phones from the audio documents, while the latter suggests a set of keywords to be checked in the rest of the modules with the retrieved phonetic information.

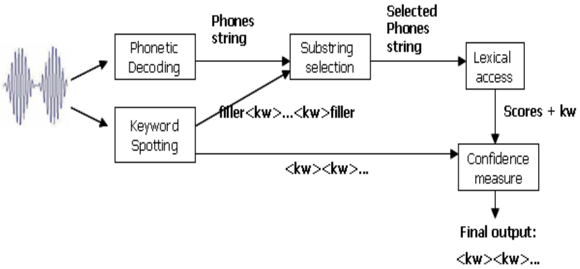

Fig. 1. Keyword spotting tool architecture

The phonetic decoding module retrieves the sequence of phones from the AM. The output of this process is a continuous stream from the 50 AMs (47 AMs plus the three types of silence). The keyword spotting module retrieves a sequence of keywords and filler models from the audio content. The filler models absorb the out-of-vocabulary words in the content. The keywords are represented as concatenations of phonetic units, so no special training data is needed to model them. A pseudo n-gram inspired in Kim's proposal [7] has been introduced as a language model, in order to investigate the performance of the system according to the frequencies (as unigrams) of both keywords and filler models. A pseudo 2-gram was used in our experiments.

The substring selection module checks the time correlation between the keywords proposed by the keyword spotting module, and the phonemes string proposed by the phonetic decoding module. The lexical access module proposes a keyword, among the phonemes string retrieved by the previous module, which best matches this string. A score is computed for the proposed keyword, based on a previously trained set of (substitution, deletion and insertion) errors, occurring in the sequence of phones retrieved by the phonetic decoding module. Finally, the confidence measure module computes an assessment of the predicted certainty on the spotted keywords in order to reduce the false acceptance rate, by discarding the results below a certain threshold α.

The output of the keyword spotting system is an XML document that reflects each match in the audio documents containing information such as the keyword, the document, the time interval where the keyword was found in the document, and a precision measure that reflects the probability of correctness of the identification. For instance, an entry for the keyword “andalucia” may look as follows:

```
<match>
  <kw>andalucia</kw>
  <doc>E3233.wav</doc>
  <tIni>2220.000000</tIni>
  <tEnd>2710.000000</tEnd>
  <prec>-3828.5259</prec>
</match>
```

This output is provided to the annotation module, as described in the previous section. It must be noted that the keyword-spotting system is language-independent. The acoustic models and the database are the only elements to change from a language to another.

5. User interface

The simplicity of the query model in traditional keyword-based IR technologies allows an extremely simple query interface, essentially consisting of a text field and a button. This cannot be said of the query model on which our retrieval system is based, which requires appropriate user interfaces (UIs) that properly isolate the user from the complexities of ontology-based representations. The research presented in the previous sections has been complemented with the development of a search UI, which exploits the semantic richness of the underlying ontologies upon which the search system is built.

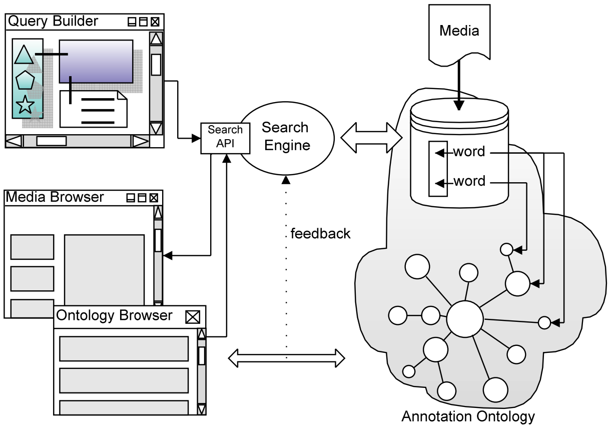

Fig. 2. High-level system architecture

The proposed UI comprises three different components, as shown on the left side of figure 2: a query builder, where the user interactively enters her/his queries; a media browser, which presents the audio content returned in response to the queries, along with the metadata associated to the content; and an ontology browser, to visualize the knowledge available in the associated domain KBs.

The query builder includes a form-based interface where the user can create structured queries by selecting ontology classes, setting conditions on their properties, etc. Alternatively, the user can enter keyword-based queries, in such a way the system builds a formal query consisting of the concepts associated to the keywords, with no conditions.

Query results are shown in the media browser. This interface provides support for browsing the returned audio pieces, showing the associated metadata as well. The views are generated by a general-purpose RDF rendering tool [4]. The displayed multimedia metadata includes a part based on the Dublin Core schema for editorial metadata (title, date, author, etc.), and a part for content-based metadata based on an MPEG-7 ontology [5]. The latter is used to model the relevant segments because of which the content was selected for a query. In addition, a specialized audiovisual view is presented, where the user can play the audio results (and associated video, if any), and interact with it through a clickable version of the audio track.

The browser allows the user to view the list of spotted words that appear in each returned audio document, from which two additional actions are supported. First, it is possible to click on any keyword in order to perform a new query for all content in the database where that concept appears. Second, the keyword view is enriched with links to the ontology. Each word that is represented by an ontology concept is linked to a description of that concept, which is shown by the knowledge browser.

In the knowledge browser, the user can navigate through the knowledge structures of ontologies and domain KBs. The tool is based on the same RDF rendering module as was mentioned in the media browser. Using the three UI components combined, the user can, for instance, find statements made by a soccer coach, then click on a team mentioned by the coach in his speech, browse the available information about the team in the KB (e.g. players, results), select a player, retrieve audio clips by or about him, etc. In this dual browsing experience, the user can thus navigate through audiovisual content in the media browser, and through the underlying semantic models, using the knowledge browser in a complementary way.

6. Experimental work

The techniques described in the previous sections have been tested in several experiments. Empirical results of the ontology-based IR model on a text corpus of considerable scale can be found in [1]. The current experiments with the speech processing framework are so far based on a corpus of limited scale at the semantic layer, whereby obtaining formal integrated results is still work in progress at the time of this writing. Nonetheless, early results at the speech processing level are reported next.

The experiments have been based on the Spanish Albayzin collection [10], comprising two corpora, each consisting of a training set and a test set. The first corpus is composed of phonetically balanced sentences, from which we have selected a training set of 4,800 sentences pronounced by 164 different speakers. The second corpus is composed of geographic sentences, from which we have selected a training set of 4,400 sentences by 88 different speakers, and a test corpus of 2,400 sentences by 48 speakers. To test the audio annotation tool performance, 80 keywords (the most representative set, appearing 1672 times, excluding stop words) were selected from the test set of the geographic corpus.

Two performance measures have been used to evaluate our keyword spotting tool: the Detection Rate (DR), which is defined as the correctly spotted keywords over the total spotted keywords, and the False Alarm Rate (FAR), defined as the incorrectly spotted keywords over the total of correctly and incorrectly spotted keywords. Table 1 shows the results for the different filler models in terms of DR and FAR and also shows the threshold α used in the confidence measure module:

Filler Model	DR	FAR	α
AM	68.0%	12.8%	-0.1
PM	76.3%	12.2%	0
BM	77.9%	14.6%	0.07
APM	71.3%	14.9%	0.15
PM+ BM	**88.0%**	**16.4%**	

Table 1. Keyword spotting tool performance

It can be seen that the PM and BM filler models result in better rates, so the output achieved with the PM and BM filler models was merged (PM+BM), producing the best rate for the whole audio annotation process. We are currently extending the corpus in order to test the whole retrieval cycle. For this purpose, a small ontology on Spanish geography has been defined for the geographic audio corpus, with concepts such as cities, rivers, mountains, etc. The ontology currently includes 37 classes, 22 properties and 366 instances.

7. Conclusion

We have presented a new approach for semantic search over audio contents, which combines ontology-based models from the Semantic Web field, with Speech Recognition techniques. The semantic IR model is an adaptation of the classic vector-space model, where domain ontologies are used in place of the keyword-based indices. A novel keyword-spotting tool, based on a hybrid word-phoneme architecture has been integrated in the system. Initial experiments using the Albayzin audio database and a Spanish geographic domain ontology have been carried out, showing positive results. The rates include 88% of correctly generated annotations and 16.4% of incorrectly identified ones during a completely automatic annotation process.

8. Acknowledgements

This research was supported by the Spanish Ministry of Science and Education (TIN2005-06885).

9. References

[1] Castells, P., Fernández, M., and Vallet, D., An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval, *IEEE Transactions on Knowledge and Data Engineering* 19(2), 2007, pp. 261-272.

[2] Cuayahuitl, H., and Serridge, B., "Out-of-vocabulary Word Modelling and Rejection for Spanish Keyword Spotting Systems", *2nd Mexican International Conference on Artificial Intelligence (MICAI 2002)*, Mérida, Mexico, 2002.

[3] Deerwester, S. et al, "Indexing by Latent Semantic Analysis", *JASIS* 41(6), 1990, pp. 391-407.

[4] García, R., and Gil, R., "Improving Human–Semantic Web Interaction: The Rhizomer Experience", *3rd Italian Semantic Web Workshop (SWAP'06)*, 2006, pp. 57-64.

[5] García., R., Celma, O., "Semantic Integration and Retrieval of Multimedia Metadata", 5th *Knowledge Markup and Semantic Annotation Workshop*, 2006, pp. 69-80.

[6] Jeanrenaude, P. et al, "Phonetic-based wordspotter: various configurations and application to event spotting", *European Conference on Speech Communication and Technology (Eurospeech 1993)*, Berlin, Germany, 1993.

[7] Kim, J. et al, "A Keyword Spotting Approach based on Pseudo N-gram Language Model", *9th Conf. on Speech and Computer (SPECOM 2004),* Patras, Greece, 2004.

[8] Lleida, E. et al, "Out of vocabulary word modeling and rejection for keyword spotting", *European Conference on Speech Communication and Technology (Eurospeech 1993)*, Berlin, Germany, 1993.

[9] Maedche, A. et al, "SEmantic portAL: The SEAL Approach" In Fensel, D. et al (eds.), *Spinning the Semantic Web*. MIT Press, Cambridge London, 2003, pp. 317-359.

[10] Moreno, A., and Guirao, J. M., "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation", In Kawaguchi et al (eds.), *Spoken Language Corpus and Linguistic Informatics*, 2006, pp. 199-218.

[11] Popov, B., et al, "KIM - A Semantic Platform for Information Extraction and Retrieval", *Journal of Natural Language Engineering*, 10(3-4), 2004, pp. 375-392.

[12] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[13] Vorhees, E., "Query expansion using lexical semantic relations", *17th ACM Conf. on Research and Development in Information Retrieval (SIGIR 1994*). Dublin, Ireland, 1994.