Audio Indexing on a Medical Video Database: the AVISON Project

Grégory Senay
LIA/CERI
University of Avignon
Avignon, France

Stanislas Oger
LIA/CERI
University of Avignon
Avignon, France

Raphaël Rubino
LIA/CERI
University of Avignon
Avignon, France

Georges Linares
LIA/CERI
University of Avignon
Avignon, France

Thomas Parent
IRCAD
Strasbourg, France

Abstract—This paper presents an overview of our research conducted in the context of the AVISON project which aims to develop a platform for indexing surgery videos of the Institute of Research Against Digestive Cancer. The platform is intended to provide a friendly query-based access to the videos database of IRCAD institute, that is dedicated to the training of international surgeons. A text-based indexing system is used for querying the videos where the textual contents are obtained with an automatic speech recognition system. The paper presents the new approaches that we proposed for dealing with these highly specialised data in an automatic manner. We present new approaches for obtaining low-cost training corpus, for automatically adapting the automatic speech recognition system, for allowing multilingual querying of videos and, finally, for filtering documents that could affect the database quality due to transcription errors.

I. INTRODUCTION

AVISON is an ANR-funded project that aims to develop a platform for indexing surgery videos of the Institute of Research Against Digestive Cancer (IRCAD) database which is composed of multilingual videos for the surgeon learning. The database contains about 3000 hours of video records and is augmented by about 1500 hours each year. From the user point of view, looking for information in such huge databases is difficult due to the lack of structuring information. In this perspective, many researches have been conducted to offer by-content access to large multimedia collections, especially in the fields of text categorisation and information retrieval.

The indexing of spoken contents usually relies on two components. The first one is based on an automatic speech recognition (ASR) system, that is supposed to produce a text transcription of the speech signal. The second component is the search engine, that considers the automatic transcriptions as classical text sources, on which text-based information retrieval techniques can be applied to estimate the relevance of the documents regarding the user request.

The effectiveness of the overall process relies on the accuracy of the ASR system. Unfortunately, the nature of media, the kind of documents (interviews, lectures, etc.) or the recording conditions dramatically affect the ASR difficulty. In these situations, transcription systems suffer from a lack of robustness, particularly in specialised fields, where word error rates are frequently greater than 30%. This difficulty of speech recognition in specialised domains is mainly due to the design of ASR systems, that rely on statistical language models whose estimation requires large and relevant training sets, that are usually unavailable on specialised domains.

The AVISON project addresses three scientific issues. First, we have to find a cost effective way for training ASR on low-resourced domains. Then, the domain addressed by the AVISON project, the new medical technologies, is in constant evolution (new techniques, scientific discovers, etc.) and the audio indexing systems have to deal with the content changes (new words, new expressions, etc.). Finally, the database should be accessed as widely as possible, regardless of the language of the users, which involves IR user queries in various languages.

The research presented in this paper are organised in three axes.

1) Automatic adaptation of the ASR system (new words learning and language model adaptation)
2) Multilingual IR querying: translation of user queries for retrieving multilingual documents
3) User centred approaches for Indexation of automatic transcriptions: semi-automatic systems and self-diagnostic measures

Figure 1 presents the processing chain of the AVISON project which begins by the video capture and finishes by the storage in the database. This paper is organised as follows: Section II outlines a new approach for adapting the ASR system to the evolution of scientific discoveries and new surgical techniques. Section III describes an automatic bilingual lexicon extraction enabling to translate medical terms. In Section IV, a semi-automatic method which helps a transcriber to produce corpus is presented. The last section shows a measure which attempts to determine the erroneous documents before storing in the indexed database.

II. AUTOMATIC SPEECH RECOGNITION ADAPTATION

As mentioned before, the content of the documents to be transcribed will evolve in the time with scientific discoveries and new techniques. New words will be introduced and the ASR system should deal with them. The ASR system is classically trained a priori on a large closed corpus and does not follow the evolution of the documents to transcribe. This mismatch between the ASR system and the document content causes transcription errors. Given that the purpose of the ASR
Many works have focused on the lexical coverage problem in the field of ASR. Authors generally propose to periodically adapt the ASR lexicon with new documents manually or automatically gathered. This approach is not dynamic enough and require a constant effort of document gathering.

Meanwhile, the number of textual documents indexed by search engines has grown exponentially. In addition to providing a huge volume of textual data, the Web is a dynamic resource: new data is added and old content is updated continuously. This feature makes the Web a great resource for adapting the LM and more specifically for the task of lexicon augmenting.

The approach that we propose is composed of four stages:  
1) Transcribing the documents with the initial LM  
2) Detecting OOV words in these transcriptions  
3) Extracting words around the OOV words for querying Web search engines and retrieving documents that may contain the targeted OOV words.  
4) Inserting the new words discovered in the documents in the LM lexicon and transcribing a second time the documents

On the AVISON test corpus, composed of 4 hours of manually transcribed documents, the OOV-rate (the number of OOV words in the documents divided by the total number of words) is of about 6%. Using this technique, about 30% of the OOV words where successfully recovered in the automatic transcriptions. This has also reduced the word error rate (WER) by 1% in absolute value. The WER is the classical measurement of the quality of an automatic transcription. It is the ratio between the number erroneous words in the transcription and the number of words in the reference transcription. Further details of this approach can be found in [1].

B. Language Model Probabilities Adaptation

The quality of LM depends on the size and quality of the corpus used for learning these models. Linguistic coverage cannot be exhaustive with any closed corpus, especially when particular domains are concerned. Considering the growth of the Web, researchers naturally tried to use this information source for estimating LM probabilities. In most of the cases, this boils down to collecting domain-specific documents from the Web, and then estimating classical LMs on these documents by counting seen word sequences for estimating probabilities [2], etc. Several research reports show that, generally, probabilistic LMs estimated from the Web are less costly to obtain, but at the same time of a lower quality than LMs learned from closed corpora, mainly because the statistical distributions of the word sequences on the Web are not reliable [3].

Nevertheless, the Web is quite exhaustive, and the existence of a word sequence on the Web can constitute a relevant information that should be integrated in the LMs. Thus, we reconsidered the information obtained from the Web: rather than approximating LM probabilities, we introduced a Web-based possibilistic measure [4] that takes account of the absence of word sequences on the Web. We proposed several

Fig. 1. Process chain of the AVISON project
The estimate of Web-based possibilities is achieved by using Formula 1:

\[ \pi_n(W) = \frac{|W_n \cap \text{Web}_n| + \alpha \cdot |W_n \setminus \text{Web}_n| \cdot \pi_{n-1}(W)}{|W_n|} \]  

where \( W \) is a sequence of \( n \) or more words, \( W_n \) is the set of word sequences of size \( n \) in \( W \), \( \text{Web}_n \) is the set of word sequences of size \( n \) on the Web, \( \setminus \) is the set subtraction operator and \( 0 \leq \alpha \leq 1 \) is the back-off coefficient empirically estimated. The terminal condition for the recursion is \( \pi_0(W) = 0 \). Simply, given the word sequences of size \( n \) of an ASR hypothesis \( W \), this is the rate of word sequences of size \( n \) that are present on the web.

In the ASR process, the measure is combined with the initial LM in a log-linear manner and produces a new LM score. This measure is thus estimated dynamically and benefits automatically of new Web data.

The Web-based possibilistic approach allowed for an absolute ASR WER reduction of 2.9% on the AVISON domain-specific corpus (from 33.3% to 30.4% WER). This approach is fully described in [5] and [6].

III. BILINGUAL LEXICON EXTRACTION

An interesting challenge in the AVISON project is the multilingual accessibility of the indexed data. Each multimedia data is already available in the source language. Searching data is therefore possible only with source language keywords. For instance, a video in the English can be retrieved with keywords in English.

Our aim is to build multilingual relations on domain specific terms, in order to make possible a multilingual data retrieval system. Basically, the specialised vocabulary has to be linked with its translations in different target languages. The resulting relations can be seen as a multilingual specialised lexicon. This lexicon is then used to match target language keywords with the source language indexed terms. For the AVISON project, domain specific terms are the cornerstone to information retrieval. We focus our task on building a bilingual medical lexicon from French to English.

Building bilingual lexicon reaches the highest accuracy when using parallel corpora, which are pairs of translated texts. However, domain specific parallel corpora are relatively rare resources, that is why the Natural Language Processing community tends to use a forthcoming bilingual resource for bilingual lexicon extraction: bilingual comparable corpora. These kind of corpora are pairs of texts in two different languages, sharing common features without being exact translations of each other. These resources are widely available and can be gathered automatically from the World Wide Web. A well known freely available comparable corpus is Wikipedia. It can be seen as a set of documents, related between languages with interlingual links. We use this resource in our work without taking into account the links between languages. We want to study the efficiency of bilingual lexicon extraction methods on the whole Wikipedia corpus without the document alignment steps.

A. Context-based Approach

Extracting bilingual lexicon from non-parallel corpora is an interesting task which started sixteen years ago in [7], [8]. These works leads to the observation that a term and its translation share context similarities. Based on this assumption, many researchers paid attention to terminology extraction from comparable corpora. Some of them focused on the association between a term and its context (context size, association measure, etc.) [9], others worked on contexts similarity between the source and the target languages [10], [11]. However, the majority of the latter works stands on the use of a bilingual lexicon, in order to find anchor points in the contexts to compare. This resource is usually called seed words. Some authors studied the impact of this resource, in terms of lexicon size, general or specific domain, etc. [12]. This approach using lexical contexts of terms is the first part of our work.

B. Cognate-based Approach

In order to improve the general accuracy of the system, the context-based approach can be combined with a cognate-based approach. Basically, orthographic similarities between a term and its translation are used to extract bilingual lexicon [13]. It is a popular approach because domain specific vocabulary contains a large amount of transliteration, even in unrelated languages. In our task, the possible common etymological roots between French and English medical terms is one of the main motivation for using a cognate-based approach. One of the most popular metric to retrieve cognates is the Levenshtein distance [14]. It can be seen as an edition distance between two terms, where deletion, substitution and insertion of letters are the edition features. This cognate-based approach is the second part of our work.

C. Topic-based Approach

Finally, to be able to handle polysemy and synonymy [15], we decide to explore a topic-based approach. We assume that a term and its translation share similarities among topics built on comparable corpora. The comparable corpora are modelled in a topic space, in order to represent context-vectors in different latent semantic themes. One of the most popular methods for semantic representation of a corpus is the so-called topic model. The Latent Dirichlet Allocation (LDA) [16] fits to our needs: a semantic based bag-of-words representation and unrelated dimensions (one dimension per topic). We use this topic model to filter target terms according to their position in the semantic space. Target terms which are too far from a source term are not selected as translation candidates.

D. Experiments and Results

The context, the cognate and the topic are the three views of the multi-view approach for bilingual medical terminology extraction. First studied individually, these three approaches are
then combined in order to increase the precision of the system. The final goal of our work is to provide a high confidence in the bilingual lexicon proposed by the system. The baseline we want to compare our approach to is the combination of the context-based and the cognate-based approaches.

Tests were made on 3000 bilingual medical terms to spot, extracted from the MeSH thesaurus along with their references. The context-based and the topic-based approaches were evaluated on English and French Wikipedia as comparable corpora, using a 9000 words bilingual lexicon extracted from the Heymans Institute of Pharmacology. A baseline system reaches a F-score of 32.7%, when our multi-view approach reaches 39%, with a precision of 99.3% (76.2% for the baseline). More details about the results are available in [17].

IV. INTERACTIVE DECODING

Usually, the construction of a corpus/database is a very costly task, whether for human time or for labor cost. Considering the limits of state-of-the-art, perfect transcriptions from a purely automatic system is a long-term goal. Usually, the most common method consists of a manual correction of the transcript provided by an ASR system when errors are detected [18]. But this method still remains costly, especially when the transcription is shoddy (more than 30% of error is relatively common in specialised fields). Starting from this idea, we propose an Interactive Decoding (ID) technique where human is integrated into the decoding process to improve transcription quality. This technique is an iterative process where computer and human help together to decode the transcript.

First of all, ID starts to a normal decoding pass. Then, a self-diagnostic process suggests to the human where are the better areas of the transcription to be corrected. As the requested correction is achieved, a new decoding pass is performed by integrating the manual corrections. This step is supposed to improve the quality of the transcription by diffusing the correction to the nearby words. While the transcription can be corrected (or when the human decide to stop the process), this step-by-step process iterates.

To determine where are the inconsistency areas in the transcription, we propose three approaches to drive the human. The first method is an ID with a correction from the Left to the Right, as a standard method to correct. The second method, named Graph Density, uses directly information about the decoding process: when ASR system hesitates between a large number of different choices, it largely develops the search graph. This strategy asks for manual corrections on these areas, that correspond to the largest parts of the search graph. The last method consists in correcting semantic outliers, that correspond to the words that are out of the semantic context of the sentence.

A. Experiments and results

All experiments are conducted by using Speeral, the LIA-broadcast news system in the framework of the ESTER evaluation campaign [19]. Our ASR system performs a 32.6% WER on the test part of the ESTER corpus. The baseline (named Human only) is a left-right correction without ID.

Experiments showed that for all of the cases, ID improves significantly the correction process: post-decoding takes benefits from partial corrections, and the overall transcription process is strongly improved by the interactivity decoding scheme.

The efficiency of driving strategies depends on the initial WER of the transcription that have to be corrected. For slightly wrong documents (WER below 40%), the Left-Right correction with ID outperforms slightly the Human Only method. On highly erroneous sentences (WERs equal and above 40%), outliers-based ID yields the better performance, significantly outperforming the Left-Right-ID. The detailed results can be found in [20].

V. SEGMENT CONFIDENT MEASURE

This section presents a semantic confidence measure that aims to predict the indexability of a automatic transcription for a task of Spoken Document Retrieval (SDR). This task corresponds to an usage scenario where the user searches for contents corresponding to a written request. In such an Information Retrieval (IR) system, a search engine operates on an index database, that is built from the transcribed spoken contents. The global IR system performance is dramatically impacted by the transcription errors. This section describes a technique that aims to evaluate the transcription quality for audio indexing. This metric is named indexability. We now present what is the indexability measure and how to predict this indexability of a transcription segment.

Errors in a transcribed sentence potentially impact all search results. Consequently, each one needs a full run of SDR evaluation. The indexability measure is computed accordingly in 3 steps: (1) the targeted speech document is automatically transcribed by the ASR system, (2) for each test-query, search is performed on the whole speech database by using correct transcriptions for all segments, except for s which is automatically transcribed, (3) the resulting ranks are compared to the ones obtained by searching the full reference transcription set. Finally, indexability $\text{Idx}(s)$ of the segment s is obtained by computing the F-measure on the top-20 ranked segments, relatively to the top-20 ranking reference (i.e. the ranking on the correct transcriptions). This algorithm estimates the individual impact of the targeted segment transcription on the global SDR process, knowing the targeted ranking.

Now, we present a method which aims to predict the impact of the recognition error on the indexation process. The Prediction of indexability is computed in two steps: First, confidence measures ([21]) are extracted from the ASR system, its combines acoustic, linguistic and graph-features (which is an analysis of the number of alternative paths). Nevertheless, meaningful words are decisive for indexation. That is the reason why in the second step, semantic compactness index $\text{Sci}$ is computed. More precisely, the $\text{Sci}$ is based on a local detection of the good meaning of the context. Each context is viewed as a bag-of-words. A large corpus is used as a
reference in order to compare each reference context to the tested documents (Wikipedia in our experiment). The figure V presents the results of document classification according to the indexability threshold, by using the predicted indexability based on confidence measure (CM), semantic compactness index (SCI) and the combination of CM and SCI (CM+SCI).

Results demonstrate that the combined approach CM+SCI improves the indexability rate than CM or SCI alone (more details in [22]). In most of the cases, combination yields the best result, it enables to classify documents more than 70%, whereas CM and SCI are 10% worse. Semantic information is a relevant feature for the detection of indexable or unindexable documents.

VI. CONCLUSION

The different studies presented in this paper allow to improve the quality of the IRCAD database in terms of navigability and multilingual access. All videos presented there are about digestive cancers for the training of the surgeons. A part of the studies deals with a dynamic use of the Web which yields good performances in the specialised field of surgery and beyond this one in general field. We proposed new approaches mainly based on open corpus such as the Web or Wikipedia for improving the translation and the indexing of the videos. The first study deals with a new automatic LM adaptation scheme based on Web data: the lexicon and the LM probabilities are dynamically updated from the Web. The bilingual lexicon extraction enables the surgeons to retrieve medical data through a multilingual information retrieval system. The last way we followed concerns user-centred approaches for audio indexing. We focused on interactive system for content extraction, and on self-diagnostic methods for indexing and search processes.

We work now on the full translation of the videos, by testing probabilistic/possibilistic Web-based models that demonstrated their efficiency on ASR task.

REFERENCES