Cultural Heritage Information Retrieval by Metadata

Ivana Tanasijević and Gordana Pavlović-Lažetić

Faculty of Mathematics, University of Belgrade
{ivana,gordana}@matf.bg.ac.rs

Abstract. Information retrieval is a problem that deals with the retrieval of documents that potentially meet some criteria. The criteria can be specified in different ways and always refer to the contents of the file. The motivation for building the cultural heritage information retrieval system was a multimedia database of the intangible cultural heritage of the Balkans and the need to facilitate its manipulation. The database mostly consisted of audio material in the form of interviews and their associated text protocols which describe the characteristics and content of the audio material. A system will be presented that is designed for retrieval of unstructured protocols by their metadata. The system also provides a graphical interface for users to select the search criteria and a preview of those documents whose contents match the criteria. The system has greatly facilitated the process of obtaining documents by content that matches the query.

Keywords: Information Retrieval, Information Extraction, Metadata, Natural Language Processing, Finite State Transducers

1 Introduction

The motivation for this research was a multimedia collection of documents about intangible cultural heritage of the Balkans and the need to facilitate its manipulation. The collection was created as a result of a ten-year research project carried out by the researchers from the Institute for Balkan Studies of the Serbian Academy of Sciences and Arts. The aim of their research was to study the characteristics of the language dialects still cherished in some parts of the Balkans although they are bound to disappear or to be modified with time. As every language is a living thing, such changes are inevitable over time. This multimedia collection consists of vast number of audio and video documents, mostly in form of interview, text documents in form of protocols, scanned manuscripts, photos, publications and books. The protocols relate to audio and video documents and contain various information about characteristics of the interview. They also explain a flow of conversations.

We present a technique for earching protocols by identifying and annotating entities corresponding to the requested characteristics. Since protocols are connected with various multimedia materials from the collection (by means of the

system presented in [5]), by retrieving them we can retrieve those other related multimedia materials.

Section 2 presents related work in the field of information extraction and information retrieval implemented by finite state transducers. Section 3 is a description of the multimedia collection that we processed in this research with emphasis on its structure and importance. In section 4 we pay attention to the structure of the protocols, which are our main concern in this paper. We also highlight some of the problems that need to be solved. Methodology for solving those problems is given in section 5. Section 6 is a precise technical overview of our approach to solving the problems by techniques of natural language processing. Section 7 presents the architecture of the system that we built. The following section 8 demonstrates how extracted information is presented to a user and what are the criteria by which documents can be retrieved. The results are systematized in section 9. In the last section conclusion of accomplishments related to the initial problem, is given.

2 Related Work

Named entity recognition in Serbian is largely up to date. The problem is analyzed and elaborated in [9], from the point of view of its rich morphological system and syntactic constraints. Special attention is given to the recognition of proper names. In [2] a system for named entity recognition and tagging in Serbian is presented that relies on large-scale lexical resources and finite-state transducers. This system recognizes several types of name, temporal and numerical expressions. Finite-state automata are used to describe the context of named entities, thus improving the precision of recognition. The widest context was used for personal names and it included the recognition of nominal phrases describing a person's position.

The opportunities and challenges of applying modern information retrieval techniques to the cultural heritage domain is duscussed in [6]. The authors discuss how and why the problem of providing access to cultural heritage can be cast naturally as an information retrieval problem and present a detailed case study of applying the modern information retrieval approach in practice within libraries, archives and museums.

3 Database of the Cultural Heritage of the Balkans

The very first task for researchers working on the cultural heritage database was to find people who would be interested in talking about different topics (so-called interlocutors). Finding interlocutors was not an easy task. In the beginning, it was necessary to find a person from the place where the research was carried out, whom interlocutors trusted and who could convincingly recommend talking to researchers. In many cases, other family members or neighbors attended the interview and often contributed to the conversation.

In addition to the interviews, which are mostly audio materials, there are recordings of different customs and rituals marking the life-cycle stages such as birth, marriage or funeral. A large part of the recordings relate to the celebration of the village customs, costume balls and fairs. Other video materials contain the recorded details that follow the topic of the conversation.

Most audio materials have protocols, which are text documents that describe the contents of the recordings and the flow of the conversation. The protocols do not have unique structure, which means that they were written in different ways and with different level of details. In some of them, the topics discussed were noted down in brief, while others have the form similar to transcripts.

4 Protocol Structure

This study analyzed the protocols associated with audio and video materials. Although they appear in different forms, most of these texts contain the information such as protocol stamps, informants, researchers, other participants in the conversation, the date and place of recording, the language, the ethnicity and religious affiliation of the informants, the places and the topics discussed, as well as other observations about the conversation. Some protocols include some of this information, while others may have all or most of it. What these protocols have in common is that they are all unstructured. There is no rule on the type of data they must contain, or on theirformat. Very often, the writing style of protocols depends on the researcher who wrote that protocol. Sometimes the information is only listed in brief without clearly formed sentences. The following is an example of the protocol:

```
148-K-RUDARE-6-BS

Rudare

Razgovor vodjen 13.07.2003. u Veliko Rudare (gornje selo)
sa Radetom Mitrovicem, 1931, rodjenim u V. Rudare. U razgovoru
ucestvuju povremeno i Slavisa cija je zena iz V. Rudare, a
ovde je u gostima i Milos Savkovic, 1932, izbeglica iz
Pestova. Na B strani se ukljucuje i  Tomanovic Dusan,
1936 iz Ibarskog Kolasina, s. Padine.
Razgovor vodila Biljana Sikimic.
Snimak traje 90 minuta.


...
```

(In English:

```
148-K-RUDARE-6-BS
```

```
Rudare

Conversation was held 13 July 2003 in Veliko Rudare (upper
village) with Rade Mitrovic, 1931, born in V. Rudare. Slavisa,
whose wife is from V. Rudare, occasionally participates in the
conversation, and here is also Milos Savkovic as a guest,
1932, a refugee from Pestovo. On side B, Tomanovic Dusan,
1936 from Ibarski Kolasin, v. Padine participates too.
Interviewed by Biljana Sikimic.
The recording lasts 90 minutes.

...
```

)

At the beginning of the protocol there is generally a variety of information about the characteristics of the interview, which we call metadata. They can be accompanied by words that can help in determining the meaning the information bears. Based on the context, the information can be further added to the set of the characteristics of the conversation. For example, the sentence: "The interview was recorded in the village of Rudare" suggests that Rudare is a location of the interview. Sometimes, a location is just mentioned without any context, for example "Rudare", in which case the meaning of the location remains unclear.

However, it is not uncommon for the information to be used in the context which is not easy to resolve systematically, i.e. by taking some programming approach, although a human is able to infer the meaning from such a nontrivial context. For example, if a role of a person mentioned is not specified, it is likely that it will be guessed from the content of the conversation in the remaining text. The roles of a person may be: informant, researcher or someone who only attended the interview or was the subject of the story. So, in the following description:

```
U razgovoru ucestvuje Smiljana Lakic, rodjena u Ripnju, udata
u Subotici. Majka joj se bavila preradom vune. Milan Lakic,
poljoprivrednik, doselio se kao izbeglica za vreme rata. Cerke
Ivana Lakic i Milena Pavlovic, obe udate, Ivana ima dvoje dece.
```

(In English:

```
Smiljana Lakic, born in Ripanj, married in Subotica, participates
in the interview. Her mother used to be engaged in processing
wool. Milan Lakic, a farmer, immigrated as a refugee during the
war. Daughters Ivana Lakic and Milena Pavlovic, both married,
Ivana has two children.
```

)

it is clear that Smiljana Lakic is an interlocutor. It may be assumed that Milan Lakic is also an interlocutor, as no other context is given for his mentioning. However, it is not clear whether her daughters Ivana and Milena are also interlocutors, or just persons mentioned.

5 Methodology

A number of protocols of about 100,000 words were randomly selected and their content was analyzed. Categories of information that are important for interviews were identified and extracted. The categories extracted and discussed in this paper include:

- stamps,
- persons,
- dates,
- years,
- languages,
- ethnicities and
- religious affiliation.

Persons were classified into three categories:

- informants (or interviewees, or interlocutors),
- researchers (or interviewers) and
- others.

We then identified and listed phrase structures that form the context of the above-mentioned categories of information; this allows identification and tagging those information. For example, if some information is to be categorized as "language", it is necessary for the adjective denoting affiliation (e.g. "Serbian", "Bunjevac") to be accompanied by anidentifier which will denote a language (e.g. "speak" or "language"). Otherwise it does not necessarily refers to the language (e.g. "Serbian customs").

We processed the dates that were written out in full, the day in Arabic numbers, the month in letters, Roman or Arabic numbers and the year using two or four digits (e.g. 23.06.2003. or 23 June 2003). The category "years" contains the year of birth or the age of a person (e.g. "in the year of 1943" or "at the age of 82"). It also contains the years of historical events (e.g. "in 1999").

The information from the category "location" appears in different meanings. It may be referred to as the location where the conversation was recorded as well as the location interviewees talk about. For example, the location can indicate the place of birth, the place where the interviewees moved or where they lived all throughout their lives or spent time working. The locations very often occur in a specific context, thus it is not difficult to recognize them in the text. For example, phrases that are related to a location would be "in the village", "from the place", "a small town", "in the area of", "municipality", "born in", "married in", "moved to" etc.

The most demanding task was to identify and classify persons. We identified word phrases and expressions that closely denote roles of persons. A person who is an informant is usually placed after the word phrases such as "conversation with", "informant", "informants:", "the meeting was attended by" etc. A person who is a researcher could be found in the context such as "researcher", "researcher is", "conversation held by", "discussion was led by", "recorded by" etc. In such situations, when the context is specified directly, it is not difficult to classify the person. However, very often persons are mentioned following each other, or they are mixed with other types of information, so their role is not clearly stated and has to be inferred from a substantial part of the preceding text. From the previous example of a protocol (interview with Smiljana Lakic), it is concluded that Milan Lakic is also informant since the preceding text was about informants, although there is no immediate context which would indicate that.

It was also noticed that all the information about persons is grouped by their role, first the information about the informants, and then the information about the researchers, or vice versa. In most protocols this information is placed at the very beginning and is followed by the contents of the conversation. Therefore, protocols can generally be divided into three types of sections in terms of the role of persons. In this way, persons who appear in these sections will be accordingly tagged. Those who are not classified in the first two groups will be classified as others.

6 Application of Unitex

In this study we used a word-processing environment Unitex [7] [1] for parsing and processing large amounts of text with the help of electronic dictionaries and local grammars written using finite automata. DELAS and DELAF are dictionaries of simple words that are associated with grammatical and inflectional information [1]. The grammatical information is mainly of morphological type and corresponds to gender, number, case, tense and person. The entries in DELAS dictionaries have the following structure:

[word]. [formal description]

where the word is a lemma or the canonical form of a set of lexical units, while the formal description is a record consisting of letters and numbers and describing the attributes of these lexical units. For example, in the Serbian language a record for the word "admire" could be:

diviti.V552+Imperf+It+Ref

describing the verb as being imperfective (Imperf), intransitive (It), and reflexive (Ref) [4].

Rules for inflecting words (lemmas) from the DELAS dictionary may be described by finite state automata. Applied to a lemma from the DELAS dictionary, they produce all its inflectional forms, which constitute DELAF, the

[1] `http://http://www-igm.univ-mlv.fr/~unitex`

dictionary of inflected forms. An example of the record for "dvorane" ("halls") from DELAF would be:

dvorane, dvorana.N600:fs2q;fp1q:fp4q:fp5q

describing the word as the inflective form of the noun "dvorana" ("hall") from the inflective class 600.

A compound word is the word whose meaning cannot be inferred from its constituents. For example, a record from the dictionary of inflected compound word forms, DELACF, for "domova culture"("House of Culture" in genitive plural) might be the following:

domova kulture, dom kulture.N+Org+Comp:p2qm

describing the word as the inflective form of the noun "dom kulture", an organization (Org) and a compound (Comp).

Based on electronic dictionaries and finite state transducers, the Unitex tool is able to process the input text and insert specific labels into the output text according to specified rules. We used Serbian electronic dictionary that was developed by the Group for language technology, Faculty of Mathematics, University of Belgrade [3]. This valuable resource has been successfully used in many applications, e.g., [8]. For the purpose of this research, a library of finite-state transducers was developed with several categories depending on the information they identify. Each of these transducers identifies word phrases (one or more words) in the text that may be marked by inserting XML tags, and thus indicating specific information of interest. We developed finite state transducers for labeling stamps, persons, roles of persons, locations, dates, years, languages, ethnicities and religions. For most of these categories, the corresponding (available) dictionaries of Serbian were also used.

Transducers that tag words from the classes language, ethnicity or religious affiliation, use lists of lemmas of words that indicate a particular class of information. For example, for the class language lemmas are "Serbian", "Bulgarian", and the like, while religion affiliation lemmas include words such as "Christian", "Muslim" and the like. Then, using the dictionary of inflected formsand based on the developed transducers, Unitex system tags all the occurrences of specified words in all their inflectional forms that appear in the text. An example of the finite state transducer that recognizes the language is given in Fig. 1.

For categories such as persons and locations, dictionaries of proper names and toponyms are used for tagging those named entities. In the first pass, all persons and locations that were recognized by the corresponding dictionaries were tagged. For example, if `<First>` and `<Last>` denote the first and the last name, `<M>` denotes capital middle initial, `<PRE>` denotes a word that begins with a capital letter, and `<E>` denotes an empty word, then the following expression:

```
[<First>+<Last>] [<M>+<E>] [<First>+<Last>] [<Last>+<PRE>+<E>]
```

recognizes an occurrence of a person in the text by first name, last name, middle initial, or second last name or nickname. The expression allows the last word to be a word that is not in the dictionary, but it must begin with a capital letter.

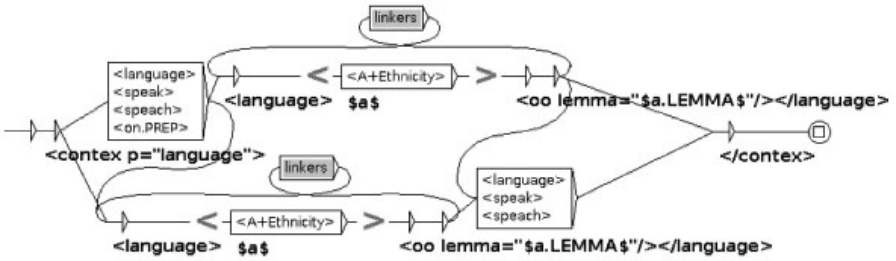

Fig. 1: Finite state transducer that recognizes entities from the class language

However, if two or more words from the previous expression are not listed in the dictionary, then it is necessary to recognize the context that clearly indicates that the occurrence of such words relates to a person. In this way the corresponding transducer could recognize an entity that is not in the dictionary. For example, in the context

```
[interview with <PRE> <PRE>]
```

the `<PRE> <PRE>` will be marked as an entity from the class persons, where the word `<PRE>` has not been recognized as a proper name by the dictionary.

The same rule is applied to identify locations. If a word begins with a capital letter and is not already recognized by a dictionary as a toponym, and if it appears after the context of the location (e.g. "in the village of Brstica"), then it will be recognized as a location by the expression:

```
[in the village of <PRE>]
```

The word recognized with `<PRE>` will be marked as an entity from the class locations. An example of a part of the finite state transducer which marks named entities by using context is given in Fig. 2.

Transducers developed were applied to protocols from the training set described in Section 5. These protocols are comprised of about 100,000 words. The obtained results were analyzed and transducers were improved by using the acquired knowledge in several iterations to increase the number of correctly marked entities and reduce the number of incorrectly marked ones. After several iterations, transducers were applied to all six hundred and eleven protocols, which consist of about 600,000 words. This is the number of protocols that are currently available in the multimedia collection of documents of cultural heritage in the Balkans. The results are shown in Table 1.

7 Information Retrieval

Text with embedded XML tags around recognized information is further loaded into the native XML database. Various queries written in the XQuery language

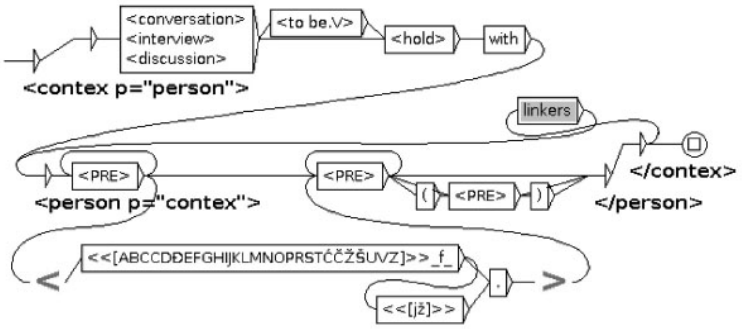

Fig. 2: Part of the finite state transducer that marks persons by using context

Table 1: The number of recognized terms by classes, based on 611 protocols

Categories	Different basic forms	Occurrences
Stamps	710	776
Informators	670	1334
Researchers	107	719
Other persons	720	1554
Locations	879	4171
Dates	398	802
Years	199	321
Languages	8	287
Ethinicities	28	2698
Religions	12	327

are then applied by means of the developed Java application to get corresponding documents or to retrieve marked information.

The system overview is presented to the user as a list of all the information which appear in the protocols, sorted by the processed categories (e.g. all informants, all places). Thus, the user has access to the information about names of all the informants and researchers who were found in the texts, as well as information from other categories. The lists are formed and presented after user authentication.

Fig. 3 shows the panel with lists of all the information found in the text. In the upper part of the panel, information is classified and listed in two rows. First row consists of the following categories: stamps, informants, researchers, other persons mentioned in the text and locations, while the second row includes the information from the categories such as dates, years, languages, ethnicities and religious affiliation. Based on this survey a user can get an informative overview of the characteristics of the protocols as a whole.

Fig. 3: Preview of extracted information and selection of search criteria

Subsequently, a user can choose one or more terms from the lists, and execute the search procedure based on the criterion that is formed as a union of the selected terms. This time, the application returns two types of texts for each of the protocols in which the information of interest is found: all the extracted information, grouped and systematized and unchanged text of protocol. The preview for the query "find all the documents about Bunjevacs" is shown in Fig. 4.

Examples of queries that can be set in this way are:

"find all the documents produced by a specified researcher and recorded on specified dates",

"find all the documents with a specified informer related to a specified location",

"find all the documents about Bunjevacs".

8 Results

The method trained by studying the characteristics of the 100,000 words of text is then applied to the texts of protocols of another 100,000 words (100 protocols). Table 2 gives an overview of precision and recall by the categories of information. Precision is defined as the ratio of the number of obtained entities which are well identified and the total number of obtained entities. Recall is defined as the ratio of the number of well identified entities and the total number of relevant occurrences of the entities in the text.

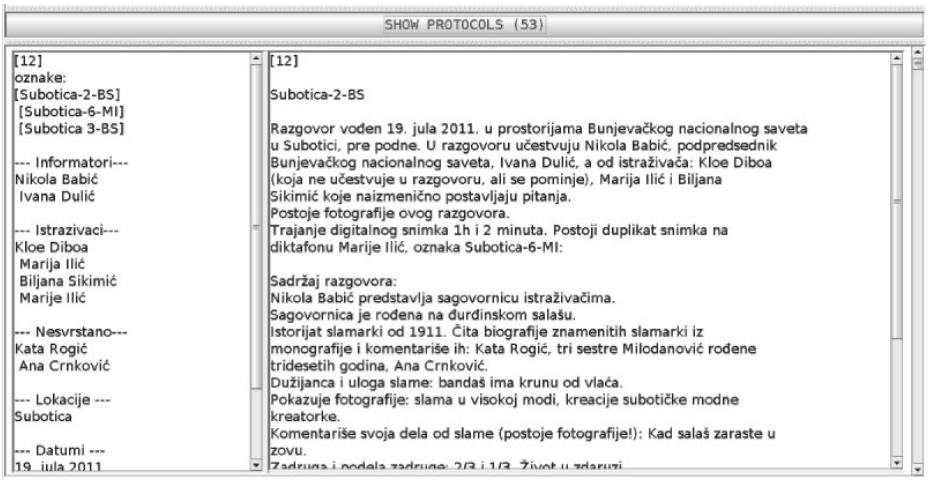

Fig. 4: Overview of documents matching the query "find all documents about Bunjevacs"

The overall precision of 0.93 and the overall recall of 0.82 may be considered as a high score. The obtained data show that we gave advantage to precision over recall.

9 Conclusion

This paper presented the main problem in dealing with large collections of multimedia documents, which is the search for the desired information. An approach to automatic processing the unstructured text documents, as one part of such a collection, is presented. The methodology for identification and labeling meta-

Table 2: Result of tagging, precision and recall counted on 100 protocols.

Categories	good tagging	wrong tagging	omitted tagging	precision	recall
Stamps	125	3	22	0.98	0.85
Informators	180	18	49	0.91	0.79
Researchers	84	5	3	0.94	0.97
Other persons	81	27	26	0.75	0.76
Locations	310	23	77	0.93	0.80
Dates	121	0	1	1.00	0.99
Years	32	3	25	0.91	0.56
Languages	24	1	7	0.96	0.77
Ethinicity	135	3	22	0.98	0.86
Religions	27	2	3	0.93	0.90

data from document protocols was introduced based on finite state transducers and electronic dictionaries. The system developed implements efficient search for metadata and thus provides for better use of the collection and its contents in other types of research.

Acknowledgments. This research was supported by the Serbian Ministry of Education and Science under the grant III47003.

References

1. Blandine Courtois and Max Silberztein. Dictionnaires électroniques du français. In *Langue Française*, n°87, Larousse, Paris, France, 1990.
2. Ivan Obradović, Cvetana Krstev, Duško Vitas and Miloš Utvić. E-dictionaries and finite-state automata for the recognition of named entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 48–56, Blois, France, july 2011. Association for Computational Linguistics.
3. Ivan Obradović, Ljubomir Popović, Duško Vitas, Cvetana Krstev and Gordana Pavlović-Lažetić. An Overview of Resources and Basic Tools for the Processing of Serbian Written Texts. *First workshop on Balkan Languages and Resources*, pages 1–8, 2003.
4. Duško Vitas, Gordana Pavlović-Lažetić and Cvetana Krstev. Towards full lexical recognition. *Text, Speech and Dialogue*, 3206:179–186, 2004.
5. Biljana Sikimić, Ivana Tanasijević, and Gordana Pavlović-Lažetić. Multimedia Database of the Cultural Heritage of the Balkans. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Language Resource and Evaluation Conference*, pages 2874–2881, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).
6. Jaap Kamps, Marijn Koolen, and Vincent de Keijzer. Information Retrieval in Cultural Heritage. *Interdisciplinary Science Reviews*, 34(2-3):268–284, 2009.
7. Sébastien Paumier. *Unitex 3.0 User Manual*, 2011. Available at `http://www-igm.univ-mlv.fr/~unitex/UnitexManual3.0.pdf`.
8. Gordana Pavlović-Lažetić, Vesna Pajić, Duško Vitas and Miloš Pajić. WebMonitoring software system: Finite state machines for monitoring the web. *Computer Science and Information Systems / ComSIS*, 10(1):1–23, 2013.
9. Duško Vitas and Gordana Pavlović-Lažetić. Extraction of Named Entities in Serbian using INTEX. In M Silberztein S Koeva, D Maurel, editor, *Formaliser les langues avec l'ordinateur: De INTEX a Nooj*, pages 281–302. Universitaires de Franche-Comte, 2007.