Deriving semantic annotations of an audiovisual program from contextual texts

Luit Gazendam[1], Véronique Malaisé[2], Guus Schreiber[2], Hennie Brugman[3]

[1] Telematics Institute, Enschedé, The Netherlands
[2] Free University, Amsterdam, The Netherlands
[3] Max Planck Institute on Psycholinguistics, Nijmegen, The Netherlands

Abstract. The aim of this paper is to explore whether indexing terms for an audiovisual program can be derived from contextual texts automatically. For this we apply natural-language processing techniques to contextual texts of two Dutch TV-programs. We use a Dutch domain thesaurus to derive possible metadata. This possible metadata is ranked by an algorithm which uses the relations of the thesaurus. We evaluate the results by comparing them to human made descriptions.

1 Introduction

The paper reports on a case study in which we used natural-language processing (NLP) techniques to derive a candidate set of indexing terms from a thesaurus, and subsequently rank these terms based on their semantic relations in the thesaurus.

The case study is carried out at the Netherlands Institute for Sound and Vision (S&V) in the course of the Choice research project. S&V archives Dutch public radio and television programs for their testimonial value (in the cultural-heritage sense) and for creation of new programs (such as documentaries). At S&V, Choice studies how semantic web and natural language processing techniques can be applied in order to keep a rapidly growing digital archive accessible[4]. With 20.000 hours of digital audio visual material each year in need of being archived, supporting cataloguers with the indexing of programs is one of the tasks of Choice. During the indexing the cataloguers consult a set of textual resources to identify the most relevant metadata for a given program. The metadata is of two types: plain text descriptions of the program's topic and a set of indexing terms. These indexing terms are selected from a thesaurus: a controlled vocabulary partially organized in a hierarchical way (according to the ISO 2788/5964 defined Broader Term / Narrower Term relationships). The thesaurus, used at Sound & Vision, is called the GTAA (Dutch acronym for Common Thesaurus Audiovisual Archives). The textual resources (henceforth to be named *context documents*) can be of several types: online TV guides, web

[4] The focus in this paper is on the application of NLP techniques. The integration with Semantic Web technology is part of future research (see section 5).

sites of the program and so on. These typically contain information like title, duration, a brief summary, etc.

The aim of this paper is to explore whether a ranked list of candidate indexing terms of an audiovisual program can be derived from contextual texts automatically. For this we apply NLP techniques to contextual texts of two TV-programs and we use the structure of the thesaurus to rank the derived metadata[5]. We use textual resources because semantic metadata extraction by movie *content* analysis is still hard. Context documents, such as TV-guide descriptions, are attractive because these provide some abstraction of the program, whereas audio transcripts of the actual program neither summarize its content nor describe explicitly what is shown in the video. This is why we conducted this experiment on context documents rather than on another genre of text. We evaluate the results by comparing them to human annotations.

This paper is organized as follows: we first present related work on Information Extraction from texts for the indexing of multimedia in section 2. Then we explain our experimental setup (section 3) and discuss the results (section 4). Section 5 reflects on this explorative study.

2 Related work

The MUMIS project [8] and the Seigo tool [6] both focus on the use of context documents to generate indexes for annotating multimedia documents. The MUMIS project deals with Dutch texts (amongst others) and merges different sources of information to improve the accuracy of the proposition generated by their extraction tools, whereas Seigo generates indexes from newspaper articles in French. They both concentrate on the sports domain (respectively football and the cycling race *Tour de France*), which is a relatively restricted domain, and take, on top of the textual corpus, a domain ontology as input. Both approaches have the goal to spot as much events as possible.

Our approach is different in the sense that we want to use the structure of a thesaurus as an external resource, and we focus on finding out the *main* topic(s) of a TV-program. This is how the indexing terms are meant to be used at Sound and Vision: they should only be associated with a program if they reflect its main topic, *i.e.* if it is relevant to retrieve the TV-program by a query composed of these indexing terms. The thesaurus structure is looser than the one of an ontology and the scope of GTAA is broad.

We do not intend to provide a complete finalized annotation of a program, instead we want to place our extraction results in a human annotation process, to help and speed it up. The ultimate decision about the relevancy of indexing terms has to be made by a human cataloguer and not by an algorithm in our approach. Therefore it is important for us to rank and order our results, so that the system can present them in an intuitive way to a human cataloguer for validation. This

[5] Currently parts of the process are applied manually as we did not yet manage to successfully implemented all pieces in a computer program

is also why we compare the results with a set of human annotations, as we try to get as close to the cataloguers's interpretation as possible.

Buitelaar and Declerck [3] define a set of linguistic analysis steps that underly Information Extraction, the part of NLP we are interested in. They define Information Extraction as the automatic extraction of "meaningful units" from semi-structured documents. In order to be semi-structured, plain text documents have to undergo the following processes:

- morphological analysis: the association of word forms to their potential part-of-speech tags (noun, adjective, verb, etc.);
- part-of-speech tagging: the selection of the relevant POS tag of the word according to its actual place and function in the sentence;
- chunking: a definition of meaningful units, as Noun Phrases, verbal groups, etc.
- dependency structure analysis: the computing of linguistic dependencies between these units, as Verb-Subject or Verb-Object relationships;
- semantic tagging: the comparison of the chunks with a list of controlled vocabulary (as terms from a thesaurus), and the search for "semantic types" in the text. The chunks can be linked to thesaurus terms if they are identical to them, their hypernyms, hyponyms or synonyms, for example. The semantic types can be location, people or organization names, and they can be spotted in the analyzed text by the application of tagging rules or grammars.

We have followed these different steps (except for the chunking, which we did not apply on top of our tokenizer, and the dependency structure analysis which we did not compute either at this first stage of experiment), and we describe them into detail along with the global experiment setup in the following section.

3 Experimental setup

3.1 Material

TV programs We used two TV-programs to test our method. The first is an item of the Dutch news bulletin "NOS journaal" on the sending of peace troops to Afghanistan. This file is referred to as *"news item Afghanistan"*. The *news item Afghanistan* was part of the late news on 29 November 2005 and addresses the political decision of sending peace troops to Afghanistan. The second is a documentary program on the fundamentalistic islam and is called: *"The Smell of Paradise"*. It was broadcasted on 11 September 2005.

Context documents Most programs are associated with multiple context documents. Expert cataloguers consider some of these to refer directly to the AV-document (*direct context documents*) and consider some others to be background information (*background context documents*). The TV-program's website text, the TV-guide's website[6] text and the technical information on the broadcasting

[6] www.omroep.nl

from an online database called Powerstation are considered as direct context documents for almost all programs. We will not take Powerstation into account in this experiment as we only focus on the extraction of:

- *keywords*, *i.e.* indexing terms describing the topic that the program is about;
- *persons*, *i.e.* indexing terms describing the people whom the program is about;
- *locations*, *i.e.* indexing terms describing the locations that the program is about or that play an important role in the program;
- *names*, *i.e.* indexing terms describing the organizations, companies, political parties etc. that the program is about;
- *makers*, *i.e.* indexing terms describing the makers of the program;
- *genre*, *i.e.* indexing terms describing the genre of the program .

The TV-programs web site text typically contains most of this required information, but it can also contain a lot of information not referring to the TV-programs content: hyperlinks to a forum, the online video, the broadcasters web site etc.

3.2 Applying natural language processing techniques

Our processing of the data with NLP techniques consist of three main parts: lemmatization, semantic tagging and ranking of the keywords.

POS tagging and lemmatization of the corpus with TreeTagger Kraaij and Pohlmann show that stemming improves the recall of an information retrieval system [5]. The use of the Part-of-Speech tag during stemming makes it more precise and therefore we assume that lemmatization will give us even better results. Lemmatization can determine which lemma a wordform is associated to (for example, *flies* can have as lemma the noun *fly* or the verb *to fly*). We choose to use TreeTagger [9]: a combined POS-tagger and lemmatizer. We use the Dutch Celex lexicon [1] as input for TreeTagger and train it with a corpus built from the Alpino Treebank, which is a tagged Dutch corpus [10]. We lemmatize and POS tag our context documents with TreeTagger before we process them with GATE: a General Architecture for Text Engineering [4]. We adjust GATE so it can cope with the TreeTagged input.

Sematic tagging of context documents with GATE We analyze these context documents using GATE, which we chose for its modularity [7]. We use GATE's Information Extraction module called Annie. We adjusted Annie so it can handle Dutch texts. In Annie we also incorporated a set of domain specific gazetteers (word lists) which we made from the GTAA. We wrote a couple of grammars to recognize named entities and domain specific notions like the starting time, the end time and the makers of a program.

Ranking algorithm for keywords The GTAA is a general thesaurus with multiple facets: subjects, genres, persons, makers, names and locations. Only the subject facet, which contains the keywords, is explicitly structured. The terms in the subject facet are related to others via the Related Term, Broader Term and Narrower Term relations. The types of information we are looking for (keywords, persons, locations, names, makers and genre) are very closely related to the different facets of the GTAA.

For the keywords (terms describing the subject), we can, on the top of the number of occurrences, also use the relations of the GTAA between the terms to get a ranking. Terms which relate to a lot of other terms found in our texts can semantically be more important than terms which we found more often but without any relations to others. A classic example of ranking based on relations is Google [2]. Instead of using the links to other pages, we use the Broader Term, Narrower Term and Related Term relations of the GTAA to connect two found terms. If one of these relation exists between two found terms we say that a relation of distance 1 exists[7]. We also check if an intermediate term connects two terms. These connections via intermediate terms are defined as relations of distance 2. We do not make any distinction in the type of relations. In figure 1 a distance 1 relation connects the terms *soldiers* and *prisoners of war*[8], which we represent by a plain line. A distance 2 relation exists between terms *prison* and *prisoner's of war*, which we represent by a dotted line. In this distance 2 relation the term *prison* is the intermediate term. A group of terms connected via these relations constitutes a semantic cluster.

We have the idea that these clusters capture the key subjects of the AV-program. We came up with the following algorithm to translate the clustered terms to a ranked list:

- Step 1. We select the keywords with both a distance 1 and a distance 2 relation. We then order these keywords based on their number of occurrences, putting the most frequent on top of the list.
- Step 2. We select the remaining keywords with a distance 2 relation to keywords found during Step 1. We order these keywords based on their number of occurrences and add them to the list.
- Step 3. We select the remaining keywords with a relation. We order these keywords based on their number of occurrences and add them to the list.
- Step 4. We order the remaining keywords based on their number of occurrences and add them to the list.

This second step might seem odd. The reason why we didn't choose in step 2 to rank higher the keywords with a distance 1 relation to a keyword found during step 1 is because it showed in practice that the keywords that had a distance 1 relation to a keyword found during step 1 were often outliers. The keywords with a relation 2 to these keywords found during Step 1 were less likely to be an

[7] We do not make a distinction between Broader Term, Narrower Term and Related Term relations in this experiment

[8] All the terms we mention in this paper are translated from Dutch to English out of consideration for our readers

outlier so we choose not to use the distance 1 relation in step 2. Another thing which is important here is the fact that in their working practice cataloguers are only allowed to attach a few keywords. When we also consider that the intended usage of the ranked list is to support cataloguers in their annotation task, it becomes clear that it is not useful to burden them with a very long list. In practice this will probably mean that we cut of the keywords below rank 5. Usually these first five ranks are found during step 1 and step 2. This makes steps 3 and 4 less important.

3.3 Evaluation

We evaluate the results of our automatic extractions by comparing them to two types of human annotation. The *expert description* is made or revised by expert cataloguers. We have this annotation for both TV-programs. The *preliminary descriptions* is made by cataloguers but was not revised by expert cataloguers. These have the status of preliminary description and were made by people from two different public broadcasters (NOS and EO) and people from the news and non-news TV-section of Sound and Vision. This set of descriptions gives insight of the diversity of human annotations and with it, the spread of reasonable annotations. We have these preliminary description only on the news item Afghanistan.

keywords	*N*	persons	*N*	names	*N*	locations	*N*	genres	*N*	makers	*N*
preliminary descriptions											
peace troops	6	Henk Jan Ormel	8	NATO	7	Afghanistan	9	the news	7	Maurice Piek	9
militairy operations	5	Jeroen Overbeek	1	CDA	3	The Netherlands	6	Topicalities	5	Jeroen Overbeek	6
armed forces	3			Pentagon	2	The United States	3	Newsbulletin	2		
government policy	2			Tweede Kamer	2	The Hague	2				
soldiers	2			Taliban	1	Kandahar	2				
				Dutch armed forces	1	Kabul	2				
				VS	1	Washington	1				
expert description											
peace troops		Henk Jan Ormel		NATO		The Netherlands		the news		Maurice Piek	
militairy operations				Dutch armed forces		Afghanistan				Jeroen Overbeek	
				The United States							

Table 1. A summary of human annotations: the preliminary and expert descriptions of the *News item Afghanistan*

The human annotations on the *news item Afghanistan* are shown in table 1. The first column on the left shows the keywords that were selected by cataloguers. The column to its right shows the number of cataloguers (total number of cataloguers is nine) that used the keyword. The expert description show no numbers. In the location column, we can see that the term *Afghanistan* was selected by all cataloguers. For the keywords in the preliminary description we see that the terms *peace troops* and *military operations* are the most frequent. These two also appeared in the expert description. When we compare the two kinds of descriptions, we can see that the preliminary description contains the set of terms selected by the expert cataloguers plus others. The human annotations for

The Smell of Paradise are displayed in table 2. This annotation was made by a professional cataloguers and checked by an expert cataloguer during the normal work process at Sound & Vision.

keywords	persons	names	locations	genres	makers
religious conflicts	Mohammed	Taliban	Chechnya	documentary	Marcin Mamon
violence		Mujaheddin	Afghanistan		Mariusz Pilis
islam			Kandahar		Tomasz Glowacki
fundamentalism			Dagestan		Marijn de Vries
extremism			Qatar		Petra Goedings
jihad			Doha		Marije Meerman
wars			Russia		
indepedence fight					
muslims					

Table 2. Human annotations for *The Smell of Paradise*. This description is checked by an expert cataloguer.

4 Results

In this section we present the results of our experiments. First we present the non lemmatized and lemmatized results. Then we focus on the relations between the keywords of the lemmatized *news item Afghanistan* and show the ordered results. Finally we compare the results with the human annotations.

4.1 Unranked results for *keywords, persons, names, locations, genres* and *makers*

keywords	*N*	persons	*N*	names	*N*	locations	*N*	genres	makers
mission	4	Balkenende	1	CDA	2	Afghanistan	7		
soldiers	3			The	2	The Hague	2		
civil servants	1			NATO	1	Europe	1		
prisoners	1			Pentagon	1				
prisoners of war	1			VVD	1				
corpses	1			Tweede kamer	1				
ministers	1			Washington Post	1				
members of parliament	1			CIA	1				
voting	1			Dutchmen	1				
				Americans	1				
				Europe	1				
				Taliban	1				
				Friday	1				
				Tomorrow	1				

Table 3. Terms derived from non lemmatized context documents by the NLP process described in section 3.2 for the *News item Afghanistan*

The terms we derived from the non-lemmatized and lemmatized context documents of the *news item Afghanistan* are shown in tables 3 and 4. In table 4 the ordered results are already shown.

ranked keywords	rank	unranked keywords	N	persons	N	names	N	locations	N	genres	makers
government	1	mission	5	Balkenende	1	CDA	2	Afghanistan	7		
soldiers	1	government	5			NATO	1	Europe	1		
prisoners of war	2	soldiers	5			Pentagon	1				
ministers	3	agreement	2			VVD	1				
prime minister	3	prisons	1			Europe	1				
prisons	4	camps	1			Taliban	1				
civil servants	4	prisoners of war	1			CIA	1				
camps	5	civil servants	1								
voting	5	ministers	1								
democratisation	5	prime minister	1								
mission	6	voting	1								
agreement	7	democratisation	1								
Christians	8	Christians	1								
lakes	9	lakes	1								
newspapers	9	newspapers	1								
writing	9	writing	1								

Table 4. Terms derived from lemmatized context documents for the *News item Afghanistan*

We have found no matches for *genres* or *makers* in either setup. In the lemmatized input we found more keywords and less *persons*, *names* and *locations*. This can be easily explained by the fact that these named entities only have one exact match whereas keywords can have multiple word forms which all reduce to one lemma. If a word corresponding to a named entity is lemmatized, it will not be recognized anymore. In the lemmatized web site text of *The Smell of Paradise* for example, the Tunisian president Zine Abidine Ben Ali is not recognized anymore because it was lemmatized to Zine Abidine zijn Ali[9]. The grammar which tries to spot person names was triggered by the combination *Ben Ali* which is not present anymore. For the *Smell of Paradise* the lemmatization of the context documents also made us find more *keywords* and less *locations*, *names* and *persons*. We did find a *genre* and *makers* for this program. The results are shown in the appendix in table 5.

4.2 Relational map and ordered results

We used the GTAA structure to find relations between the derived keywords. A visualization of the relations between the keywords for the lemmatized *News item Afghanistan* is shown in figure 1. Prisoners of war and soldiers have a distance 1 relation. A distance 2 relation is visualized by the dotted line between voting and democratization. This relation implies the existence of an intermediate term, which is elections in this case.

[9] *Ben* is an inflection of the Dutch verb *zijn*, which means *to be*.

We used the relations and the number of keyword occurrences to make a ranking among them. The results are shown in tables 4 and 5. For the *News item Afghanistan* top ranked terms are government, soldiers, prisoners of war, minister and prime minister. For the *Smell of Paradise* we have: 1. islam, 2. jihad, 3. muslim, 3. radio and 3. broadcaster as the five top ranking terms.

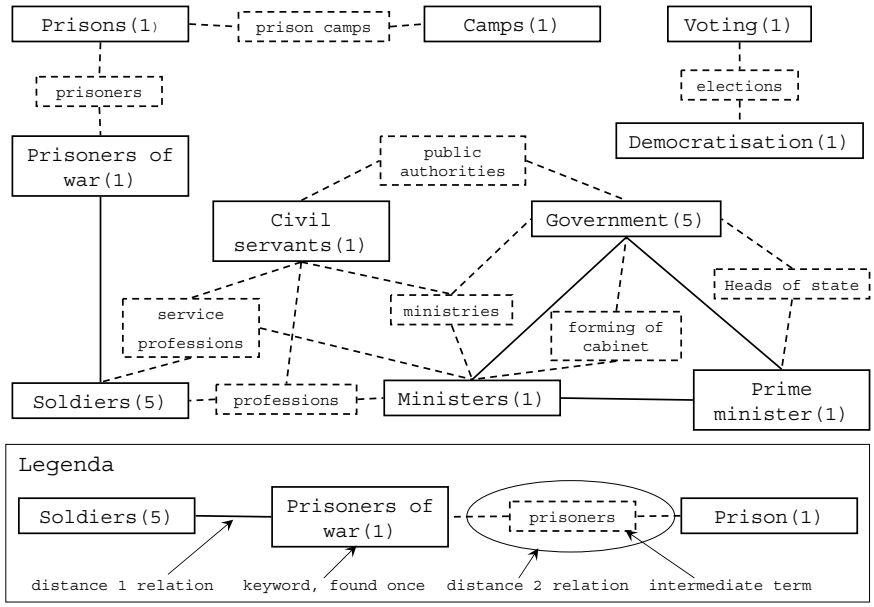

Fig. 1. Relations between keywords found for the *News item Afghanistan* with use of lemmatization

4.3 Comparing the results with the human annotations

Now we need a way to analyze the results of our method. A useful way to study the list of derived terms is by comparing the list to the human annotations. When comparing the list of derived terms with the human annotations we can encounter three different situations which de denote as *true positives*, *false positives* and *false negatives*. A *true positive* is when a derived term also appears in the human annotations. One can state that our techniques did derive a term which should have been derived. A *false positive* is when a derived term does not appear in the human annotations. A *false negative* is when a term from the human annotations does not appear in our list of derived terms.

When we label the results in this manner however, we see that some of the false positives are not that wrong at all. For these we introduce an extra label called *semi positives*, which is based on our (subjective) interpretation. A *semi positive* is a term which was derived from context documents and which did not appear in the human annotations, but which, in our opinion, does describe the content of the TV-program quite well.

In the following subsections we will label the outcome of the experiment according to the labels *true positive*, *semi positive*, *false positive* and *false negative*. For this we compare the results for *News item Afghanistan* as given in tables 1 and 4. For the *Smell of Paradise* we compare results as given in tables 2 and 5.

True positives For the *News item Afghanistan* we correctly derived the keyword soldiers, the names CDA, NATO, Pentagon, Taliban and location Afghanistan.

For *The Smell of Paradise* we correctly derived the keywords islam, jihad, muslims, the locations Chechnya, Afghanistan, Qatar and makers Mariusz Pilis, Marcin Manon and Petra Goedings. These terms both appeared in our list with derived terms and in the human annotations.

Semi positives In our opinion, the following derived terms, which did not appear in the human annotations, are semantically closely related to the topic of the *News item Afghanistan*: the keywords government and ministers. The TV-program is about the political decision making of sending peace troops to Afghanistan, so attaching these to the TV-program does make sense.

We derived the following terms that are semantically closely related to the topic of the *The Smell of Paradise*: the keyword conflict, the person Mullah Omar, the name Allah, the locations Caucasus, Islamabad, Kashmir and Pakistan and the genre informative. The keyword religious conflict appeared in human annotations, so the keyword conflict is only slightly off. These keywords have a distance 1 relation. The TV-program was for an important part about the person Mullah Omar, but the program was not solely about him. The Caucasus is the mountains between Chechnya and Dagestan where the program makers filmed the independence fighters. Both Chechnya and Dagestan appear in the human annotation. The use of a country and a mountain range is a bit less logical then using two countries, but still the use of Caucasus is good.

False positives For the *News item Afghanistan* we will name a few of the derived semantically unrelated terms: the keyword prisons, camps, weather, mission, agreement, Christians, lakes, newspaper and writing, the name Europe and location Europe. Some of these just do not reflect the key topic of the program (prisons) and some are the result of wrong lemmatization (camp). Others are the result of homonymy (mission) as the GTAA keyword mission means catholic mission and not military mission. The name Camp was lemmatized to camp, which made our grammar unable to label the phrase *minister Camp* as a person.

For *The Smell of Paradise* we also can distinguish different types of errors on the term derivation. The keywords radio, broadcaster, biography, weather, forum,

smell are found as a result of: wrong lemmatization(weather), non meaningful information of the web site such as links to a forum and background information such as biographies, domain specific noise as Radio and broadcaster which are encountered in both types of context documents and multiple occurrences in the GTAA of a term without choosing the most likely role. Petra is correctly discovered as a part of a person name in the context document, but it also appears in the GTAA as a location and a name and as a result also gets discovered as a name and location. At the moment from one piece of text multiple terms can be derived. It is probably better to only derive the most likely term.

False negatives For the *News item Afghanistan* we did not derive the keywords peace troops, military operations, the names Dutch armed forces, Tweede Kamer, location The Netherlands, The United States, genre the news and makers Maurice Piek and Jeroen Overbeek. The name Tweede Kamer appears in the text but is lemmatized and subsequently not recognized. The term peace troops did not appear in the context documents neither did any synonyms or hyponyms. This is the most frequent keyword cataloguers chose, so somehow this is the most evident. For The Netherlands adjectives appear in the context documents, but these are not linked to the location of The Netherlands. As the use of adjectives of nations happens a lot, we might look into the use of a repository mapping the adjectives of countries to the countries themselves. The makers do not appear in any context document and we did not come up with a strategy to derive them.

For *The Smell of Paradise* we did not derive the keywords violence, religious conflict, fundamentalism, wars and independence fights, the name mujaheddin locations Kandahar, Dagestan, Doha, Russia, genre documentary and some of the makers. For the makers it is just not clear from the context documents which makers contributed in what manner to the program. This lack of information in the context documents which is in the TV-program makes the locations Kandahar, Dagestan, Doha and genre documentary not derivable. The keywords religious conflict, fundamentalism and wars are semantically related to terms we have found in the context documents. However the distances to these terms and the fact that we only suggest terms which we found in the context documents makes these not derivable. We could investigate whether some of the found intermediate terms should also be taken into account or whether suggesting a common broader term is a good idea. This takes us to the discussion and future work.

5 Discussion and future work

We have seen in the section 3 that some keywords selected by human annotators were not retrieved by our system. We will investigate the gain that additional features would bring to our system, as the automatic linking of adjectives with country names, for example. This information could be derived from the Celex Lexicon, and would give us more matches of GTAA terms. We also want to explore other algorithms for ranking the keywords. Adding the distance 3 relations and discriminating against the types of relations may give interesting results.

Also it may be interesting to look into the possibility of ranking the intermediate terms too, *i.e.* suggesting terms that connect found terms, but that were not found in the context documents themselves.

When a term is not retrieved because it does not figure in the context documents at all (nor its synonym or hyponym), we will also investigate the gain we could obtain by taking into account:

- external knowledge, for example making the pars pro toto relationship explicit between Den Haag, derived form context documents and The Netherlands, which was chosen by the cataloguers. This could be made on the basis of information provided by a geographical ontology or thesaurus;
- existing documentary descriptions to build clusters of frequently co-occurring indexing terms, in order to suggest also these *termSets* as indexing terms when one or more term(s) from a given cluster is derived from a context document.

Further analysis of the context documents in order to filter out the domain specific noise may prove useful as well. Only selecting the body text of a website for example might filter out a lot of (semantic) noise.

We will also explore the relevancy of different genres of documents for information extraction: test the results of our techniques on subtitling, for example, as these texts would be likely to contain all of the makers' names.

In the Choice project we made a SKOS compliant RDF OWL representation of the GTAA. We want to see if we can use SW techniques to do some reasoning on the found indexing terms once we have represented them in a SW standard too.

Another issue raised by this experiment is the gap between human selection of keywords and automatic extraction. The keyword government was derived automatically, but it is not chosen by the cataloguers, although it does relate quite closely to the content of the program. We think that various phenomena are at stake here. The cataloguers do not mention information which they consider too common knowledge. The sending of peace troops to Afghanistan was discussed for a long time in the Dutch government so the mentioning of the term government is perhaps considered as too obvious. Cataloguers may think that people looking for this program just might not use these terms to retrieve the program. This last point might be the most complicated issue to deal with: how to select the relevant information content to be as efficient as possible in helping human cataloguers in the indexing process? We will have to pay a particular attention to this precise notion of relevancy.

Acknowledgements

We would like to acknowledge Mettina Veenstra and Mark van Setten for their comments on earlier versions of this paper and Lora Aroyo and Cristian Negru for their help in obtaining the preliminary descriptions from cataloguers.

References

1. R. H. Baayen, R. Piepenbrock, and L. Gulikers, *The celex lexical database*, (release 2) [cd-rom] ed., Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA:, 1995.
2. Sergey Brin and Lawrence Page, *The anatomy of a large-scale hypertextual Web search engine*, Computer Networks and ISDN Systems **30** (1998), no. 1–7, 107–117.
3. Paul Buitelaar and Thierry Declerck, *annotations for the semantic web*, vol. 1, ch. Linguisitc Annotation for the Semantic Web, pp. 93–110, IOS press, 2003.
4. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, *GATE: A framework and graphical development environment for robust NLP tools and applications*, Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
5. Wessel Kraaij and Renée Pohlmann, *Viewing stemming as recall enhancement*, Proc. of SIGIR '96, 1996, pp. 40–48.
6. Estelle Le Roux, *Extraction d'information de documents textuels associs des contenus audiovisuels*, In proceeding of the RECITAL conference (Tours), 2001, pp. 503 – 507.
7. D. Maynard, V. Tablan, and H. Cunningham, *Ne recognition without training data on a language you don't speak*, ACL Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models (2003).
8. Horacio Saggion, Hamish Cunningham, Kalina Bontcheva, Diana Maynard, Oana Hamza, and Yorick Wilks, *Multimedia indexing through multi-source and multi-language information extraction: the mumis project*, Data Knowledge Engineering **48** (2004), no. 2, 247–264.
9. Helmut Schmid, *Improvements in part-of-speech tagging with an application to german*, Proceedings of the ACL SIGDAT-Workshop, March 1995.
10. Leonoor van der Beek, Gosse Bouma, Jan Daciuk Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoa Villada, *Algorithms for linguistic processing*, ch. 5. The Alpino Dependency Treebank, NWO PIONIER Progress Report, 2002.

6 Appendix

Ranked keywords	rank	keywords	N	persons	N	names	N	locations	N	genres	N	makers	N
islam	1	smell	8	Mullah Omar	2	VPRO	5	Chechnya	5	biography	6	Mariusz Pilis	3
jihad	2	biography	6	Tony Blair	2	Towers	2	Qatar	4	informative	1	Marcin Manon	2
muslims	3	islam	5	Ahmed Shah Massoud	1	Europe	2	England	3			Petra Goedings	1
radio	3	jihad	3	Doke Romeijn	1	Allah	2	The Netherlands	3			George Brugmans	1
broadcaster	3	death	3	Frank Wiering	1	21st century	1	Afghanistan	3			Doke Romeijn	1
research	4	president	3	George Brugmans	1	Petra	1	Europe	2			Frank Wiering	1
conflict	5	interview	3	Mullah Mohammed Omar	1			Marocco	1				
photos	5	muslims	2					Caucasus	1				
video	5	internet	2					Islamabad	1				
promo	5	research	2					Kashmir	1				
legislation	5	weather	2					Tunesia	1				
biography	6	summer	2					Pakistan	1				
death	7	broadcaster	2					Petra	1				
interview	7	actions	2										
democracy	8	people	2										
actions	8	sculpture	2										
weather	8	spider	2										
summer	8	directing	2										
forum	8	radio	2										
book	8	book	2										
directing	8	producer	2										
producer	8	democracy	2										
borders	9	forum	2										
reincarnation	9	photos	1										
fire	9	fire	1										
villa	9	borders	1										
housing	9	aurora	1										
smell	10	villa	1										
president	11	shop	1										
internet	12	housing	1										
people	13	name	1										
sculpture	14	reincarnation	1										
spider	15	conflict	1										
bee	16	legislation	1										
water scource	17	bee	1										
aurora	18	water scource	1										
shop	19	video	1										
name	20	promo	1										

Table 5. Terms derived from lemmatized context documents for *The Smell of Paradise*, displaying both the unordered and ranked keywords