Extraction of novelty concepts from TV broadcasts with longitudinal user experiments

Jouni Sarvanko
University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
jouni.sarvanko@ee.oulu.fi

Mika Rautiainen
University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
mika.rautiainen@ee.oulu.fi

Arto Heikkinen
University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
arto.heikkinen@ee.oulu.fi

Mika Ylianttila
University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
mika.ylianttila@oulu.fi

Jukka Riekki
University of Oulu
P.O.BOX 4500
FIN-90014 UNIVERSITY OF OULU, Finland
jukka.riekki@ee.oulu.fi

ABSTRACT

Recent popularity of online catch-up TV services has facilitated time-shifted TV viewing. However, contemporary services do not utilize the rich information available in broadcast TV content streams. Richer program descriptions and summaries help on-demand viewers to employ new information seeking behaviour to find interesting content. Techniques for content-based analysis of broadcast TV streams aim to improve access to relevant archived TV content and assist in efficient on-demand viewing. In this paper we introduce a methodology that extracts novelty concept words from Finnish broadcast TV stream. The methodology is employed in an online content analysis system, which executes near real-time analysis and indexing of seven free-to-air DVB TV channels. The methodology uses machine learning and statistical data mining techniques to extract descriptive novelty concepts automatically from TV program subtitles. Extracted concepts are further used to summarize and access program content in end-user services that facilitate search and browsing of archived TV content. We show results from user logs of nearly 3 000 sessions to demonstrate how novelty words have been used in our prototype services.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information System**]: Video; H.3.3 [**Information Search and Retrieval**]: [Retrieval models]; H.3.1 [**Content Analysis and Indexing**]: [Indexing methods, Linguistic processing]; H.2.8 [**Database Applications**]: [Data mining]; I.2.7 [**Natural Language Processing**]: [Text analysis]

Academic MindTrek 2013, October 1-4, 2013, Tampere, FINLAND.

General Terms

Algorithms, Experimentation, Measurement

Keywords

video analysis, novelty concept extraction, online TV

1. INTRODUCTION

Conventional video summarization methods focus on the audiovisual content characterization. Volumes of data in time-continuous digital video offers rich data source for alternative summarization techniques. Additionally, video broadcasts are often supplemented with subtitles, which provide very semantic but underutilized data source for semantic video access. A popular way of summarizing text and web document collections is using word or tag clouds, which are typically based on frequency or popularity of a set of user assigned tags or document key words. Several methods exist for key word extraction [7][6][12].

We have adapted word cloud model from web text domain to automatic analysis and summarization of linear DVB video content based on subtitle data. Instead of using popular tags or most frequent words to create word clouds, we introduce a method for extracting concept words from broadcast subtitles based on their statistical novelty. We use it to summarize broadcast TV programs in online end-user services that have been developed for browsing and seeking information from archived TV content.

Novelty detection is the identification of new or unknown data or signal that a machine learning system is not aware of during training [5]. Another definition was given by Soboroff & Harman for NIST TREC Novelty track [9]: "the task was to highlight sentences containing relevant and new information in a short, topical document stream. This is analogous to highlighting key parts of a document for another person to read, and this kind of output can be useful as input to a summarization system." Based on these definitions, we describe our novelty detection task as following: highlight new and relevant information from broadcast TV stream to summarize topical content for novel online services.

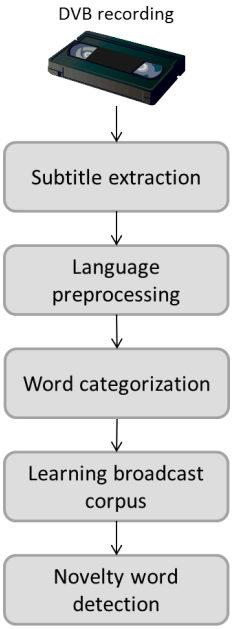

Figure 1: Novelty word detection sequence

This paper describes methodology and applications that propose a solution to the task for DVB broadcasts. In section 2 we give a detailed look into the methodology and in section 3 we present two end-user applications that have utilized the methodology. Section 4 describes user experiments and section 5 contains discussion and conclusions.

2. METHODOLOGY FOR EXTRACTING NOVELTY CONCEPTS FROM BROADCAST TV

The process contains five phases seen in Figure 1. First the videos are captured and their subtitles are extracted from bitmaps or fetched from teletext. Then the language is preprocessed to fix and prepare the text extracts for further analysis. The words are divided into categories before they are collected into broadcast corpora and used in novelty word detection.

2.1 Video capturing and subtitle extraction

We capture DVB broadcasted TV programs and extract DVB subtitles or teletext if available. Also, if both DVB subtitles and teletext subtitles are unavailable, our system detects embedded subtitles from video frames. We use image morphology and histogram analysis to preprocess bitmap text. For optical character recognition (OCR) we use GOCR [2] and Tesseract [1] open source software. For embedded subtitles, we post-process detected subtitle segments using error correction based on multi-sample word recognition and difference image verification to clean srt encoded subtitles. In general, we estimate that we are able to obtain well above 90% of all subtitles in the selected TV channels. The channels use many subtitling techniques with preference in teletext and DVB subtitle formats. Embedded subtitles are common in commercial channels.

2.2 Language preprocessing

Before we could perform any analysis, we had to fix remaining OCR mistakes and lemmatize and group the words of the TV programs. One of the most important clean-up tasks was to remove frequent words that did not convey any topical information in themselves e.g. conjunctions, adverbs etc. We collected these words into a black list and added on the list single words and word groups that were so neutral and commonplace that they would not specify the program in any way For example, the most common verbs and adjectives were placed on the list because they contained words like 'olla' (to be), 'sanoa' (to say), 'hyvä' (good) and 'suuri' (big) that could be used in a plenitude of different situations. The black list was used to filter out these low information words and to leave only words that could be potentially used to summarize the contents of the program.

Also, we had to implement specific logic to recover malformed words into their proper forms. For example, after subtitle extraction we had a lot of words that mistook letter 'a' for 'o' or vice versa. Some of the words had changed to other proper words (e.g. 'olla' (to be) to 'alla' (under)), which made them invisible to our algorithms, but most of the changed words were unrecognisable. These we could fix by changing the problematic letters one-by-one and by going through all the possible combination in an effort to find a recognizable word.

For some common yet complex linguistic tasks we used free programs for Finnish language processing. These are listed below.

1. Voikko [11] was used to analyse the structures of the words.
2. Sukija [10] was used to get the lemmas (dictionary forms) of words.
3. Snowball [8] was used as a fallback to find the stem (root form) of words algorithmically.

Most helpful were the two Finnish grammars (libraries) for Malaga: Voikko [11] and Sukija [10]. These used vocabularies to process words, which did not work for uncommon words e.g. foreign names. As a fallback in these situations, we used Snowball [8], which is a stemming algorithm that contains also rules for Finnish language. It produced the stem (root form) of words, which we applied to combine inflected words that Malaga libraries did not recognize.

2.3 Word categorization

We used three labels to categorize words. These are listed below.

1. Generic words: common nouns, adjectives, verbs etc.
2. Names: proper nouns
3. Abbreviations: acronyms and initialisms

Generic words are simply all the recognized and accepted words that are not names or abbreviations. Names are words that are either recognized as names by Voikko [11] or interpreted as such through additional logic. Abbreviations are all the recognized or interpreted abbreviations – or acronyms and initialisms. The more common abbreviations which are only written with small letters in Finnish

Table 1: Normative corpus size

TV programs (subtitled)	140 000
Generic words	274 000
Unique names	53 000
Abbreviations	5 000

Table 2: Variable notations for two corpora

	Corpus 1	Corpus 2
Frequency of a word	O_1	O_2
Total amount of words	N_1	N_2

were filtered out because they did not convey any interesting information about the text e.g. 'jne.' which corresponds to 'etc.' in English.

2.3.1 Special names and abbreviations

We also applied Voikko [11] and Sukija [10] to check if word candidates were recognizable Finnish words. This approach worked for generic words, but names and abbreviations contained peculiarities that the dictionaries did not know e.g. foreign names. Therefore we needed to implement some special logic to separate these words from trash.

Our approach was to stem all the unrecognized words with Snowball [8] and gather together words with similar word stem. If a group of words under a word stem formed many enough variations, the word stem was considered a proper word and was accepted. In this case it was decided that the shortest word of the group would be considered the lemma for the group. Words that did not have the required amount of variations were judged to be trash and removed.

With abbreviations, we took advantage of colons because they are used to inflect most of the abbreviations in Finnish e.g. 'USA:ssa' (in USA). This allowed us to separate the abbreviation from its suffix. Similar to name words, unrecognized abbreviations needed a few variations to be accepted.

2.4 Learning broadcast corpus

We needed a normative corpus to work as a baseline for our novelty detection. To train this corpus, we used all the programs and their words from our collection of broadcast media. The size of the corpus is seen in table 1.

The corpus was based on frequencies of the words inside the whole collection. The frequencies were normalized per program to remove the bias to longer and more word heavy programs. This we did by sorting the words of a program by their frequency into descending order and then applying linear ordinality-based normalization. We employed normalization equation 1 to calculate normalized frequency for each word.

$$f_i = \frac{n - i + 1}{n} \tag{1}$$

Here n is the amount of unique words in the program, $i = \{1, ..., n\}$ is the order of the word in the sorted list and f_i is the normalized frequency for the word. These normalized word frequencies of the programs were counted together to form total frequency scores for the words in the normative corpus.

2.4.1 Sample corpora

For the later phases of the process, we needed corpora to compare against our normative corpus. These corpora were to be the samples from which we wanted to detect novelty words. They were collected from programs within specific genres over restricted time periods. Since the data was in essence the same as was used to train the normative corpus, these sample corpora were in fact subcorpora of the normative corpus.

The used time periods were one day, one week and one month. We also created smaller genre-base corpora by further dividing these time-based subcorpora with program genres. Each genre contained a list full titles or words expected to be found from the titles of the programs that belonged to that specific genre. For example news, cooking and nature programs formed their own genres.

2.5 Novelty word extraction

Our aim was to select subcorpora from our normative corpus and see if there were any statistical differences between the two. These variations could be considered as identifying concepts for the sample corpus.

Our novelty word extraction methodology was applied and re-purposed from equations 2 and 3 in [7]. Table 2 explains the used notations. In [7] these equations were used to analyse two corpora. However, we re-purposed them to extract novelty words, i.e. words that are non-normative in a typical TV broadcast. The frequency of a word was compared against its expected value E on each corpora. The logarithmic values of these ratios were weighted with the word's frequency, which balances out the benefit for the less frequent words. Also, we were only interested in the words that were more frequent in the sample corpus. Thus, we ignored the words that were less frequent than was expected from the normative corpus.

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{2}$$

$$L = 2 * ((O_1 * log(O_1/E_1)) + (O_2 * log(O_2/E_2))) \tag{3}$$

In order to increase variety, we created a method to detect unique words that were descriptive for the subcorpora. We modified the equation 3 by removing the weight balance. This elevated the words with low frequency. In Finnish language this typically means very unique compound words. This is due to the agglutinative nature of Finnish of which a symptom is a wide use of compound words – some of which are very long and infrequent. We also removed the part of the formula that used the baseline corpus as a balancing factor. This was done to further emphasize the frequency changes in the subcorpora.

We used these two set of rules on generic word list (chapter 2.3). This gave us generic (equation 3) and special (modified equation 3) novelty word lists for the generic word

Figure 2: Catch-up TV Guide program view for a program about Ötzi the Iceman.

group. Novelty word lists for name and abbreviation lists were generated only with equation 3. For clarity all the four different novelty word lists are listed below:

1. Generic
2. Special
3. Name
4. Abbreviation

3. END-USER APPLICATIONS

We have used the novelty word lists in two end-user applications. In both cases, different novelty word lists (listed in chapter 2.5) have been combined together to get more diversity.

3.1 Catch-up TV Guide service

Our Catch-up TV Guide service or Mediaseinä (media wall) [3] is a TV guide that gives summarizations of recently broadcasted TV programs. The service supports browsing and following programs on the Web, and links to a program's catch-up web stream if it is available on the broadcaster's site. The service utilizes generic and special novelty word lists to summarize the most novel concept words in the programs. The word list forms a program specific novelty word cloud that visualizes novel topics with different weights. Novelty word cloud is also a collection of hyperlinks, that allows users to access relevant excerpts from the program. This is achieved by extracting dynamically generated picture quotes from the video content. With this design, users do not need to establish and navigate entire video streams to access relevant content. Instead they can assess the relevancy of the information before committing to viewing the stream. We support viewing the actual video stream by linking to actual video at broadcaster's site. The aim of Catch-up TV Guide is to give easy access to the most relevant content of the programs and facilitate finding new interesting TV programs for users with varying information needs.

Figure 2 shows a program box in the Catch-up TV Guide. It displays a TV documentary about Ötzi the Iceman. Starting from the top, in order from left to right, the words in the novelty word cloud are: Borreliosis, human, iceman, iceman institute, Stone Age, contamination, copper axe, hand, lactose intolerance, world, stomach, arrowhead, cavity organ, made of/containing flint, melt edge.

3.2 Novelty Cloud service

Second prototype end-user service we have developed is Novelty Cloud or Uutispilvi (news cloud) [4]. It is a service that collects the most notable topics mentioned in news broadcasts during a week or a month. It combines generic, special and name novelty word lists and generates a large novelty word cloud from them. In contrast to Catch-up TV Guide, this service displays novelty word summaries from a group of programs over a period of time instead of displaying individual programs. It can be utilized to quickly get the big picture of news events from the time period.

Figure 3 is an example word cloud from the week 36. It is the week when the news about Microsoft buying Nokia came out. The cloud is full of related words like 'Nokia', 'Elop', 'Microsoft', 'phone factory' (puhelintehdas), 'technology supplier' (teknologiatoimittaja) etc.

Users may click any of the words on the Novelty Cloud to retrieve the TV programs where the word has been present. Furthermore, users are given access to the original news broadcast at the broadcaster's site if they are still shared. With the introduced design, service allows users to skim through news broadcasts for the past weeks and months in a quick and effortless manner.

Figure 4 contains the word clouds for the week 16 and the month of April of the year 2013. These share the same time period during which a few news events gained high attention globally.

The week cloud contains most notably words like 'Boston', 'bomb strike' (pommi-isku), 'marathon' (maraton) and 'of Chechen background' (tsetsheenitaustainen), which relate to the horrid event of the April. Another major disaster, the earthquake of Lushan, was less visible and only the words 'Sichuan' and 'magnitude' (magnitudi) relate to that event. National events like pension quarrel and floods of Pyhäjoki are more in the headlines. Pension topic is mentioned with words 'pensio quarrel' (eläkerahariita), 'retirement

Figure 3: Novelty cloud from week 36 of the year 2013.

expert' (eläkeasiantuntija) and 'payment burden' (maksurasitus). The spring floods are less visible but contain many words like 'flood' (tulva), 'ice dam' (tulvapato), 'flood situations' (tulvatilanne) and a list of affected regions and rivers that link to the news about the floods (Pyhäjoki, Lapväärtinjoki, Loimijoki, Etelä-Pohjanmaa, Satakunta).

In month cloud, the global events of the week cloud have almost disappeared. Only the word Boston is left and Lushan related terms do not show any more. Instead the local events of the week have still quite strong influence. Especially the pension issue is visibly the topic of the month. Also, a single magazine program about the youth of Libya has a notable appearance in both clouds with the names 'Gaddafi', 'Benghazi', 'Abdalsalam' and 'Nabbus'. This is most likely because the topic has been out of media focus for a longer period of time. It can be justified since it brings out new information of an old topic.

4. LONGITUDINAL USER EXPERIMENTS WITH ONLINE TV SERVICES

We have captured and analysed user logs for the introduced end-user services. Catch-up TV Guide has been part of a longitudinal study to examine how semantic cues are used to access broadcast information. We analyzed user logs from one year period (2/2012-2/2013) and obtained over 2 500 user sessions that accessed the Catch-up TV Guide service. We calculated how often users clicked novelty concept words, highlighted key frame images and program titles as depicted in Figure 2. This was done to examine and compare the level of interest in linguistic word summaries, visual cues and titles of TV programs. Table 3 shows the access statistics of TV programs and relative percentages. Novelty concept words were the most popular way to preview information in TV programs. 650 sessions (26% of all sessions in Catch-up TV Guide service) accessed program content information by clicking novelty concept words, highlighted key frames or program titles. On average there were approximately one novelty word click, 0.57 highlighted key frame clicks and 0.38 titles clicked per session.

Table 3: Access statistics of TV programs in Catch-up TV Guide service

TV Program access method	Frequency	Percentage
Via novelty words	646	51%
Via key frames	373	29%
Via program titles	248	20%

Novelty Cloud end-user service has been operational since 2013. It provides weekly and monthly novelty topics from January 2012 up to current week. 410 user sessions have accessed Novelty Cloud end-user service, from which 132 sessions (32% of all sessions in Novelty Cloud service) accessed 228

Figure 4: Novelty clouds from the week 16 (left) and the month of April (right) of the year 2013.

extracted novelty concepts since January 2013. The average of accessed novelty concept words per session was 1.72. This is higher than the average accessed novelty word in the Catch-up TV Guide service.

5. DISCUSSION AND CONCLUSIONS

Result from our longitudinal user experiments indicate that proposed methodology for extracting novelty concept words was found more interesting than picture highlights or program titles. Extracted novelty words were attracting more clicks per session in the Novelty Cloud service than in the Catch-up TV Guide service. This result could be caused by the design differences. The TV Guide service supports more casual skimming of recent TV programs whereas the Novelty Cloud service encourages people to find more information about programs by clicking words. Overall use statistics show that the proposed methodologies attract user interest in both end-user services.

We have introduced a methodology to enrich semantic summarization of DVB broadcasts and demonstrated its applicability in novel online TV services with nearly 3 000 user sessions. We believe that the results introduced in this paper are relevant for the video and TV broadcast research. Our experiments show that access to TV program information can be enriched with content-based analysis of subtitles and that the extracted novelty concepts are interesting for the end users. Future work involves incorporating personalized preferences to novelty detection, applying methodology to English language, integrating with social media and experiments with second screen application concepts.

6. ACKNOWLEDGMENTS

We would like to thank National Technology Agency of Finland and Academy of Finland for funding our research and acknowledge the ITEA 2 program and ACDC and ESENS projects. A special thank is given to Aki Mikkonen for his work on solving OCR problematics.

7. REFERENCES

[1] Google. Tesseract. `https://code.google.com/p/tesseract-ocr/`. [Accessed 13 September 2013].

[2] Joerg Schulenburg. Gocr. `http://jocr.sourceforge.net/`. [Accessed 13 September 2013].

[3] Kuukkelitv.fi. KuukkeliTV - Mediaseinä. `http://www.kuukkelitv.fi/mediaseina`. [Accessed 13 September 2013].

[4] Kuukkelitv.fi. KuukkeliTV - Uutispilvi. `http://www.kuukkelitv.fi/uutispilvi`. [Accessed 13 September 2013].

[5] M. Markou and S. Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481 – 2497, 2003.

[6] Y. Matsuo and M. Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(01):157–169, 2004.

[7] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora*, CompareCorpora '00, pages 1–6, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[8] Snowball. Snowball. `http://snowball.tartarus.org/`. [Accessed 13 September 2013].

[9] I. Soboroff and D. Harman. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[10] Sukija. Sukija. `http://sourceforge.net/projects/sukija/`. [Accessed 13 September 2013].

[11] Voikko. Voikko. `http://voikko.sourceforge.net/`. [Accessed 13 September 2013].

[12] K. Zhang, H. Xu, J. Tang, and J. Li. Keyword extraction using support vector machine. In J. Yu, M. Kitsuregawa, and H. Leong, editors, *Advances in Web-Age Information Management*, volume 4016 of *Lecture Notes in Computer Science*, pages 85–96. Springer Berlin Heidelberg, 2006.