A Novel Lecture Browsing System Using Ranked Key Phrases and StreamGraphs

Martin Gropp, Elmar Nöth, and Korbinian Riedhammer

Lehrstuhl für Informatik 5 (Mustererkennung)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstr. 3, 91058 Erlangen, GERMANY
korbinian.riedhammer@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de

Abstract A growing number of universities offer recordings of lectures, seminars and talks in an online e-learning portal. However, the user is often not interested in the entire recording, but is looking for parts covering a certain topic. Usually, the user has to either watch the whole video or "zap" through the lecture and risk missing important details. We present an integrated web-based platform to help users find relevant sections within recorded lecture videos by providing them with a ranked list of key phrases. For a user-defined subset of these, a StreamGraph visualizes when important key phrases occur and how prominent they are at the given time. To come up with the best key phrase rankings, we evaluate three different key phrase ranking methods using lectures of different topics by comparing automatic with human rankings, and show that human and automatic rankings yield similar scores using Normalized Discounted Cumulative Gain (NDCG).

Keywords: key phrases, ranking, visualization, browsing, e-learning

1 Introduction

A growing number of universities offer e-learning material to both their students and, to some extent, the public. Aside from lecture slides or work sheets, many schools provide audio or video recordings of lectures, seminars and talks. Software solutions like *iTunes U*[1] or *OpenCast*[2] help with the recording, storage and organization of the data.

Most e-learning sites provide the user only with a catalog of audio and video recordings, sometimes annotated with short descriptions, tags or user comments. The documents themselves are presented as is, usually an audio or video file with play, pause, rewind and seek controls which is sufficient for a student who is interested in the whole recording.

However, the same archives are often used as supplemental material for preparing a class project or studying for an exam. In these cases, the user is interested in whether or not a certain topic or key phrase is mentioned in the recording and if so, when and in what context it occurs. As an example, consider a student refreshing a class on machine learning who is interested in regression

[1] http://www.apple.com/education/itunes-u
[2] http://www.opencastproject.org

and classification. Without the information mentioned above, he or she would either have to listen to all recordings, which is very time-consuming, or try to "zap" through all the recordings to spot some key words, hoping to catch all relevant parts of the lecture.

In this work, we first evaluate three different key phrase ranking strategies in a small user study that was part of a student thesis [3]. In a next step, we introduce an integrated interactive web-based platform which provides the user with the recording, a ranked list of important key phrases and, for a subset, a visualization of when these key phrases appear, which can be used to navigate within the recording. Together with the possibility to manually add, delete or re-rank key phrases, this integrated tool can greatly increase the utility of the recordings and contributes to making e-learning easier and more efficient.

2 Related Work and Motivation

The motivation for this work is to step away from extractive summarization of spoken language as it is limited in terms of readability and involves the risk of omitting important details. Instead, we provide the user with a tool to find all the information he or she needs within a short period of time. The user is presented with the raw audio/video data to avoid recognition errors and confusions due to utterances extracted without context.

The goal of the proposed integrated platform to is to help the user find the information he or she is looking for in the lecture. This task is closely related to (query-based) extractive summarization, which is the concatenation of salient utterances. In [11], a tool for interactive meeting summarization was presented. The user is provided with an initial set of weighted key phrases which are used to compute an extractive summary. The user can then modify the key phrases and their weights to produce summaries that contain the requested information. Although the interface was never thoroughly evaluated, follow up work [12] confirmed that well weighted key phrases can be used to compute very good extractive summaries compared to human abstracts.

Similarly, we extract and rank key phrases to provide an initial overview of the lecture, but choose a graphical representation (see Sec. 6) instead of an extractive summary. Key phrase extraction is traditionally divided in supervised (i. e., a previously trained classifier decides whether or not a phrase is salient, e. g. [6]) and unsupervised methods which do not require prior training. Recent work on meeting summarization utilize part-of-speech n-grams, lexical chains, or graph-based methods, e. g. [12, 7]). Others suggest to extract key phrases using a ranking approach [5] with Learning-to-Rank methods as found in the information retrieval community [8].

3 Data

The BASE Corpus[3] consists of 160 lectures and 40 seminars recorded in a variety of departments (video-recorded at the University of Warwick and audio-recorded at the University of Reading). It contains 1,644,942 tokens in total (lectures and seminars). Holdings are distributed across four broad disciplinary groups, each

[3] http://www2.warwick.ac.uk/fac/soc/al/research/collect/base

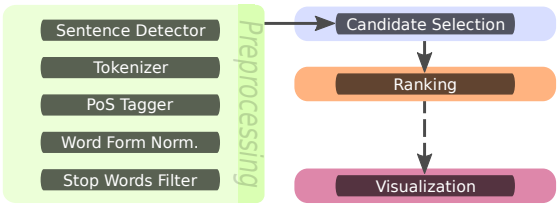

Figure 1. Key phrase extraction process.

represented by 40 lectures and 10 seminars. In this work, we focus on the "Arts and Humanities" lecture series *(ahlct)*, namely lecture *008* on Huckleberry Finn, and lecture *009*, an introduction to Assembler programming from the "Physical Sciences" group *(pslct)*.

4 Key Phrases

4.1 Candidate Selection

Fig. 1 shows an overview of the key phrase extraction process. The input data, either the output of a speech recognition system (ASR) or, as for the BASE corpus, a manual transcription, is split in chunks by the sentence detector using annotations like punctuations or pauses (transcription) or prosodic cues (ASR). The tokenizer prepares the input for the Part-of-Speech (PoS) tagger [9]. Word form normalization [10] and a stop words filter (about 900 words containing conversational speech artifacts) finalize the pre-processing.

The candidate selection is taken from [12] and can be summarized as matching PoS patterns against a regular expression allowing certain sequences of tags modeling noun phrases. Example key phrase candidates extracted from lecture *ahlct008* are *"Huckleberry Finn"*, *"Mark Twain"*, *"Tom Sawyer"*, *"American literature"*, and *"civil war"*.

4.2 Ranking

No Language Model Similar to [12], we design a heuristic ranking function which combines the frequency (n) of a candidate phrase g with its n-gram length using a weighting function w that emphasizes phrases of length 2 or 3

$$\text{f x len}(g) = n \cdot w(n_t) \quad , \quad w(x) = x \cdot e^{-\frac{1}{5}x^{3/2}} \tag{1}$$

where n_t is the number of words within the phrase with PoS tags indicating nouns, foreign words, numbers, adjectives or gerunds.

Corpus Specific Language Model A common feature in information retrieval is the term frequency (TF) multiplied by an inverse document frequency (IDF) giving a notion of how document-specific a word or phrase is. The IDF values are estimated on a representative document collection of the target domain. Here,

we consider an ideal setup where we estimate the IDF values on all lectures of a series, i. e., we get two corpus specific IDF values for the series *ahlct* and *pslct* which can be integrated in the frequency based ranking as

$$\text{tfidf x len}(g) = n \cdot \text{IDF}(g) \cdot w(n_t) \tag{2}$$

General Background Language Model A more general approach is to compare phrase occurrences to a general background language model. We compare phrase distribution probabilities estimated on the *British National Corpus* [1] using a point-wise Kullback-Leibler (KL) divergence [13]

$$\text{KL}(g) = p(g) \cdot \log_2 \frac{p(g)}{q(g)} \tag{3}$$

where $p(\cdot)$ and $q(\cdot)$ are the document and background phrase probabilities. Note that there is no heuristic correction for phrase length for the same reason as in KS.

5 Evaluation

5.1 Setup

For the evaluation of the system human raters were given transcripts of the lectures. Unlike the direct output of speech recognizers, these transcripts were first checked for recognition errors, divided into sentences and meaningful paragraphs, and had punctuation added. It can be assumed that these superficial changes, while dramatically improving readability, do not affect the way humans understand a text, and therefore have no impact on the key phrases identified by the rater.

The raters were asked to read the lecture and produce a sorted list of 20 key phrases they thought were most suitable for representing its content. The relevance of these phrases should be rated on a scale ranging from 1 (very relevant) to 6 (extraneous)[4]. Any unrated phrase was assumed to belong to the worst category and given a 6.

An analysis of the phrases selected by our algorithm showed that there were almost no good phrases that were not ranked among the best 15 by at least one of the methods. So, in order to make things easier for the raters, they were provided with a list of candidates to choose from which consisted of the best 20 phrases according to each of the employed relevance measures. This typically resulted in about 50 alphabetically sorted items. Under the assumption that all sensible candidates appear on this list, this method should have no effect on the results. However, the raters were encouraged to add more phrases from the text.

5.2 Evaluation Measure

The Normalized Discounted Cumulative Gain (NDCG) [4] rewards placing "valuable" items at the top of a retrieved list. Every phrase g_i is assigned a "gain"

[4] These correspond to the German school grading system which has turned out to produce more homogeneous results than other scales.

(the more relevant, the higher) multiplied with a discount factor based on its position in the list (the further to the back, the lower).

In order to emphasize the top ranks, we use an exponential function to map the grade assigned by the rater to a gain value:

$$\text{gain}(g_i) = 2^{(6-\text{grade}_i)/5} - 1 \tag{4}$$

The discount function suggested by Järvelin and Kekäläinen is the reciprocal logarithm, but since there is no particular reason for this exact function, we use $1/\log_2(1 + \text{pos})$ to avoid any special treatment for the first item. The base of the logarithm can be varied depending on how much the top ranks should be emphasized.

The DCG is then simply the sum over the discounted gains of all phrases

$$\text{DCG}_N = \sum_{i=1}^{N} \frac{\text{gain}(g_i)}{\log_2(1+i)} \quad , \quad \text{NDCG}_N = \frac{\text{DCG}_N}{\text{ideal DCG}_N} \tag{5}$$

which is then normalized by division by the ideal DCG (of a sorted list).

5.3 Results

The lectures *ahlct008* and *pslct009* were each evaluated by five human raters (computer science students). We are now interested in the quality of both human and automatic rankings.

To ensure a fair evaluation, we select one human rater's relevance scores to calculate the NDCG scores for the remaining human and the automatic rankings. This is repeated five times to use each rater's relevance scoring once. Finally, all human NDCG scores are averaged to single NDCG score (human). Similarly, a mean is computed for each automatic ranking method (f x len, tfidf x len and KL).

Fig. 2 and 3 show the NDCG scores for the lectures *ahlct008* and *pslct009*. The Y axis represents the evaluation measure, where 1 is the best achievable value. The X axis specifies the number of key phrases considered for the evaluation (beginning with the top ranked key phrase).

As expected, the human rankings (continuous line) receive consistently good scores which decrease with the number of key phrases considered. This makes sense as the raters strongly agree on the really important key phrases but not necessarily on the less important ones.

The general observation is that, considering a useful number of key phrases per lecture, e. g. 5 to 10, the automatic rankings are comparable to human rankings, i. e., it is hard to tell the difference between an automatic and a human ranking. Furthermore, integrating suitable language model information helps to keep the automatic ranking consistent with the human rankings. For the humanities lecture *ahlct008*, the general background language model seems to be more adequate, while the technical *pslct009* can benefit from the corpus specific information.

6 Integrated Browsing System

Although the key phrase extraction and ranking algorithms produce reliable results, just a textual representation is often not sufficient to get an overview of

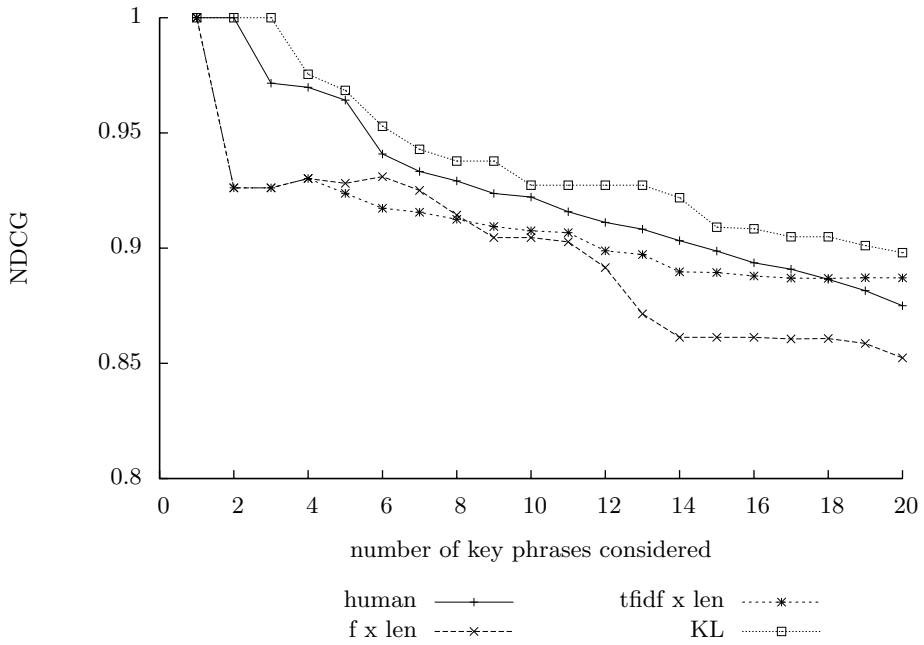

Figure 2. NDCG scores for lecture *ahlct008*.

the whole lecture. Thus, we integrate them into a browsing interface as depicted in Fig. 4: The StreamGraph [2] on the bottom left visualizes when important key phrases occur and how prominent they are at the given time by mapping a key phrase to one colored wave. This can be used to navigate within the video: by clicking on the desired position on the StreamGraph (horizontal for time, vertical for phrase and dominance), the video playback begins a few seconds before the occurrence of the requested phrase. The list on the right shows the available key phrases and controls which phrases should be included in the graph, usually about five. Furthermore, the user can remove existing or add further key phrases as desired.

Once the system is in regular use, statistics about favored, deleted and added key phrases can be collected, which are the basis for combining existing ranking methods by Learning to Rank.

7 Summary

In this work, we compared four unsupervised methods to rank automatically extracted key phrases. We conducted a small user study on lectures of different topics and could show that the best automatically ranked key phrases are of similar quality as human rankings, especially for shorter lists of key phrases.

Furthermore, we motivated and described a web-based platform for lecture video browsing. In addition to the video, we provide an automatically extracted

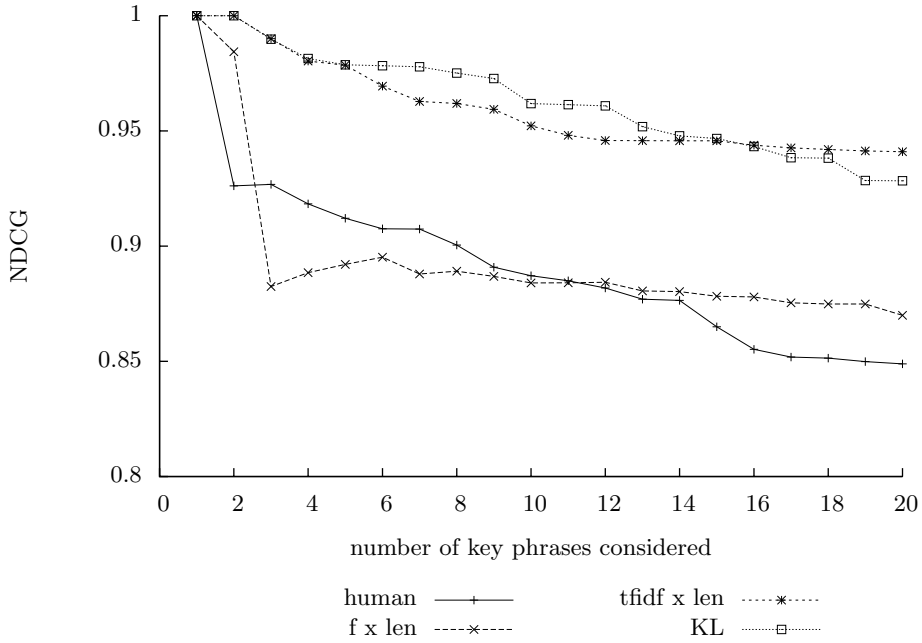

Figure 3. NDCG scores for lecture *pslct009*.

and ranked list of key phrases. A user-defined selection is displayed in a StreamGraph that visualizes when the phrases occur and how prominent they are at the given time. This allows the user to quickly find the information he or she needs and provides a more natural presentation of the data in contrast to extractive summarization.

The interface allows to collect statistics about certain user interactions, e. g. which key phrases are visualized most in the StreamGraph or which key phrases were added or deleted. These data can then help to develop better unsupervised methods or build the basis for supervised Learning-to-Rank methods.

8 Acknowledgments

Part of the transcriptions used in this study come from the British Academic Spoken English (BASE) corpus project. The corpus was developed at the Universities of Warwick and Reading under the directorship of Hilary Nesi and Paul Thompson. Corpus development was assisted by funding from BALEAP, EURALEX, the British Academy and the Arts and Humanities Research Council. Part of this work was supported by the European regional development fund (ERDF) under STMWVT grant IUK-0906-0002 in cooperation with the Medav GmbH.

References

1. Burnard, L. (ed.): Reference Guide for the British National Corpus. Research Technologies Service at Oxford University Computing Services (February 2007)

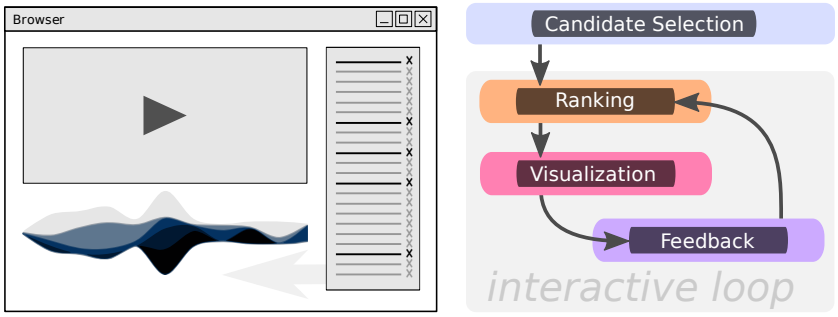

Figure 4. *Left:* A mock-up of the integrated browsing system. The video can be controlled by clicking into StreamGraph below. The list on the right shows the available and key phrases (displayed phrases in bold face). *Right:* Diagram of the user interaction with the key phrase selection, ranking and display.

2. Byron, L., Wattenberg, M.: Stacked Graphs – Geometry & Aesthetics. IEEE Transactions on Visualization and Computer Graphics (TVCG) 14(6), 1245–1252 (2008)
3. Gropp, M.: Key Phrases for the Textual and Visual Summarization of Academic Spoken Language. Studienarbeit Informatik, Dept. Informatik 5, Univ. Erlangen-Nuremberg (2010)
4. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20(4), 422–446 (2002)
5. Jiang, X., Hu, Y., Li, H.: A ranking approach to keyphrase extraction. In: Proc. ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 756–757 (2009)
6. Liu, F., Liu, F., Liu, Y.: Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In: Proc. IEEE Workshop on Spoken Language Technologies (SLT). pp. 181–184 (2008)
7. Liu, F., Pennell, D., Liu, F., Liu, Y.: Unsupervised approaches for automatic keyword extraction using meeting transcripts. In: Proc. NAACL-HLT. pp. 620–628. ACL, Stroudsburg, PA, USA (2009)
8. Liu, T.Y.: Learning to Rank for Information Retrieval, Foundations and Trends in Information Retrieval, vol. 3. now Publishers (2009)
9. Phan, X.: CRFTagger: CRF English POS Tagger (2006), `http://crftagger.sourceforge.net`
10. Porter, M.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
11. Riedhammer, K., Favre, B., Hakkani-Tür, D.: A Keyphrase Based Approach to Interactive Meeting Summarization. In: Proc. IEEE Workshop on Spoken Language Technologies (SLT). pp. 153–156 (2008)
12. Riedhammer, K., Favre, B., Hakkani-Tür, D.: Long Story Short – Global Unsupervised Models for Keyphrase Based Meeting Summarization. Speech Communication 52(10), 801–815 (2010)
13. Tomokiyo, T., Hurst, M.: A Language Model Approach to Keyphrase Extraction. In: Proc. ACL Workshop on Multiword Expressions. pp. 33–40 (2003)