Selection and Ranking of Text from Highly Imperfect Transcripts for Retrieval of Video Content

Alexander Haubold

Department of Computer Science
Columbia University
New York, NY 10027
ahaubold@cs.columbia.edu

ABSTRACT

In the domain of video content retrieval, we present an approach for selecting words and phrases from highly imperfect automatically generated transcripts. Extracted terms are ranked according to their descriptiveness and presented to the user in a multimedia browser interface. We use sense querying from the WordNet lexical database for our method of text selection and ranking. Evaluation of 679 video summarization tasks from 442 users shows that the method of ranking and emphasizing terms according to descriptiveness results in higher accuracy responses in less time compared to the baseline of no ranking.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, selection process*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems – *evaluation/methodology*.

General Terms

Measurement, Performance, Design, Experimentation.

Keywords

Speech transcripts, imperfect transcription, relevancy, ranking.

1. INTRODUCTION

Search and retrieval of video content rely heavily on cues drawn from text in the form of speech from automatic speech recognition (ASR) and writing from optical character recognition (OCR). Existing multimedia content browsers for news [1] and sports use textual cues mostly from ASR, which is reasonably accurate for edited videos with high-quality audio. ASR transcripts from unedited videos, such as lecture and presentation videos recorded in noisy environments exhibit higher word error rates (WER) of up to 75% [2]. Selection and relevancy ranking of key terms from these transcripts is difficult. Information retrieval methods, such as TF-IDF fail on the raw data, because most of the unique yet incorrectly identified terms are salient.

We focus on building text indices for unedited student presentation videos from a large undergraduate engineering course (> 150 students per term). Recorded presentations are

made available to students for self- and peer evaluation, instructor review, and archival. We use the VAST MM browser [4] to disseminate streaming videos and their visual summaries. ASR transcripts from these videos are unusable in their raw format (see Table 1). Presentation videos are on purpose candidly captured in a classroom environment without special consideration for recording quality. A standard microphone is used to capture audio and a commercial ASR software package produces approximate transcriptions. We note that building speaker models to increase accuracy is infeasible due to the large number of students.

2. TEXT SELECTION AND RANKING

We apply relevant external indices to filter raw transcripts. We have shown that course textbook indices are effective filters for lecture videos [2]. Raw transcripts from student presentation videos can best be filtered with text content from presentation slides or other course material. We first extract all text from presentation slides, including bullet points, paragraphs of prose, titles, captions, etc. Each of the resulting text blurbs is then filtered for meaningful words and phrases using the WordNet [3] lexical database. Stemming takes place only to remove plural senses, because fully stemmed terms cannot be queried in WordNet. (from hereon we use "phrases" to also denote "words").

Set A: Phrases that do not exist in the WordNet database are extracted by identifying all WordNet-resolvable terms and selecting consecutive words in-between them. Oftentimes this includes named entities and technical terms, e.g. "... met with *Dr. Topsy Krets* to discuss ...".

Set B: We separately identify all WordNet phrases in the text blurb by iterating over each word and resolving the longest possible phrase starting with this words and including successive terms. In this iteration, we do not remove stop words, which are helpful in identifying proper phrases, e.g. "Statue of Liberty".

Set C: All single terms that are not stop words are extracted into a separate set, including those which have formed phrases.

The combination of these three sets is applied as a filter to the raw imperfect transcripts. The resulting list of matching time-aligned terms serves as a searchable index into video content. We note that it is very difficult to identify complete sentences, since inaccurate automatic transcriptions not only lack punctuation, but also rarely contain grammatically valid phrases. Alternatively, the filter itself can be used as a superficial index; however, only the intersection with speech transcripts resolves phrases in specific locations and temporal recurrence in long video sequences.
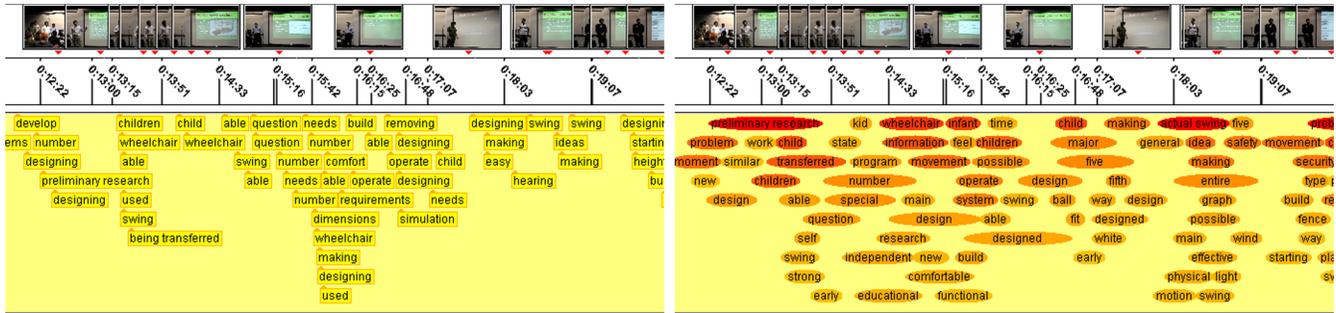
Figure 1: Video browser showing keyframes and key phrases for 8 minutes of a student presentation. Left: Text is not ranked. Right: Phrases are ranked (red=higher, orange=lower). Studies favor the interface on the right for visually selecting key concepts.

Table 1: Raw automatically generated transcript. Valid terms are highlighted in bold. Index terms are underlined. This presentation describes the design of a wheelchair swing.

... **preliminary research and design** the river begun **similar solutions** the with and then vote in the doing **next few weeks** for work and the **impact** on a long haul that we're given was **that children who** one any the **use** on **swings** of modern this lens **without being transferred** to standards went subject of why this suit bill the slaying that eight of **child** who's in los share of **would be able to swing** and **without** been **transferred** to slight and if **they're strong enough** case also be **able to** of **self** propelled ...

Table 2: Evaluation of summarization tasks with and without text ranking and pre-defined multiple choice answers, and with text ranking and articulation of response.

Accuracy	With Rank	Without Rank	w/ Rank, articulate
1 (most)	87%	67%	82%
2	10%	20%	15%
3 (least)	3%	13%	3%
Duration	41.85 sec	51.47 sec	87.98 sec

We assign to each identified phrase a numerical rank, which is used to emphasize text differently in the browser interface (Figure 1). Terms considered unique or very descriptive are ranked higher, whereas more common terms are ranked lower:

1a: If the phrase does not exist in the WordNet database (terms from Set B), then the phrase's rank is equal to its number of words * T [numWords * T]. T is a numerical value which can be adjusted to emphasize the weight of a named entity in comparison to the measures used to rank terms from other sets.

1b: Resolve the number of synonymous senses [numSenses] for WordNet phrases. Less specific phrases have higher values.

2: Using WordNet's hypernym sense querying, establish the distance of a phrase's sense to the root sense [distRoot] (e.g. a noun's root sense is "entity"). The more distant a phrase is to the root, the more specific and descriptive it is. Without performing sense disambiguation due to highly noisy text data, we use the most common word sense defined in WordNet for resolution.

3: If the phrase contains only nouns, boost the rank of the phrase, because pure noun phrases are generally highly specific. (e.g. [nounEmphasis] = 3 * number of nouns).

4: The final rank of a phrase is the linear combination, which relates numSenses inversely to numWords and distRoot:

$$\frac{numWords * distRoot}{numSenses} + nounEmphasis$$

5: The user interface separately allows the user to cluster equivalent phrases within a variable temporal vicinity via a slider, resulting in extended phrase ovals (Figure 1). When clustered, the weight of a phrase increases linearly.

3. INTERFACE AND EVALUATION

We have compared two user interfaces (Figure 1), one without and one with ranking of text. In both interfaces, phrases are automatically time aligned and placed in a panel with a greedy algorithm, filling space towards the top first. In the interface featuring text ranking, color and vertical position are used to emphasize text of varying weight. Similar phrases are grouped into single blips within a user-controlled time interval. This interactive setting expands blips horizontally to mark the duration over which the represented text is used in the video.

We have evaluated our method of phrase selection and ranking in a user study with 313 students over two semesters. Students were presented several tasks related to search and retrieval, including summarization of unfamiliar presentations. In this task, students are presented only with the filtered phrases for a presentation, which they must use to formulate an idea about its content to then select a fitting description from a multiple-choice menu. We measure performance by the duration of a task and the accuracy of the response. Each interface was used in one of two consecutive semesters. Results show that with increasing accuracy of responses, required time per task decreases (Table 2).

We have also performed a separate user study with 129 students and 240 summarization tasks, in which students were required to articulate their response. We note that the duration of this task increased to 88 seconds due to the lack of preset responses, but accuracy remained at a high level (Table 2).

4. REFERENCES

[1] Campbell, M., Haubold, A., Ebadollahi, S., Naphade, M.R., Natsev, P., Smith, J.R., Tesic, J., and Xie, L. IBM Research TRECVID-2006 Video Retrieval System. *Proc. TRECVID 2006 Workshop*. NIST Special Publications, 2006.

[2] Haubold, A., and Kender, J.R. Analysis and Visualization of Index Words from Audio Transcripts of Instructional Videos. *MCBAR '04*. IEEE Press, New York, NY, 2004, 570-573.

[3] http://wordnet.princeton.edu

[4] http://www.aquaphoenix.com/research/vastmm