Presentation Video Retrieval using Automatically Recovered Slide and Spoken Text

Matthew Cooper

FX Palo Alto Laboratory
Palo Alto, CA 94034 USA
cooper@fxpal.com

ABSTRACT

Video is becoming a prevalent medium for e-learning. Lecture videos contain text information in both the visual and aural channels: the presentation slides and lecturer's speech. This paper examines the relative utility of automatically recovered text from these sources for lecture video retrieval. To extract the visual information, we apply video content analysis to detect slides and optical character recognition to obtain their text. Automatic speech recognition is used similarly to extract spoken text from the recorded audio. We perform controlled experiments with manually created ground truth for both the slide and spoken text from more than 60 hours of lecture video. We compare the automatically extracted slide and spoken text in terms of accuracy relative to ground truth, overlap with one another, and utility for video retrieval. Results reveal that automatically recovered slide text and spoken text contain different content with varying error profiles. Experiments demonstrate that automatically extracted slide text enables higher precision video retrieval than automatically recovered spoken text.

1. INTRODUCTION

Presentation video is a rapidly growing genre of Internet distributed content due to its increasing use in education. Efficiently directing consumers to video lecture content of interest remains a challenging problem. Current video retrieval systems rely heavily on manually created text metadata due to the "semantic gap" between content-based features and text-based content descriptions.

Presentation video is uniquely suited to automatic indexing for retrieval. Often, presentations are delivered with the aid of slides that express the author's topical structuring of the content. Shots in which an individual slide appears or is discussed correspond to natural units for temporal video segmentation. Slides contain text describing the video content that is not available in other genres. The spoken text of presentations typically complements the slide text, but is the product of a combination of carefully authored scripts and spontaneous improvisation. Spoken text is more abundant, but can be less distinctive and descriptive in comparison to slide text.

Automatically recovering slide text from presentation videos remains difficult. Advances in capture technology have resulted in higher quality presentation videos. As capture quality improves, slide text is more reliably recovered via optical character recognition (OCR). TalkMiner is a lecture video search engine that currently indexes over 30,000 videos for text-based search. TalkMiner automatically selects a set of keyframes containing slides to represent each video. OCR is used to extract the slide text appearing in the videos' keyframes for indexing and search. Details of the content analysis and indexing for this system are found in Adcock, *et al.*[1] Currently, the search indexes in TalkMiner rely heavily on automatically recovered *slide* text. A recent focus is incorporating automatically recovered *spoken* text from the videos' audio streams. Extracting spoken information from presentation video via automatic speech recognition (ASR) is also challenging. The canonical video recording setup is optimized for visual information capture by focusing the camera on the slides from the back of the lecture room. This default choice for camera operators can improve the accuracy of slide text extraction. In contrast, audio quality can be poor in this capture scenario, particularly in the absence of additional microphones nearer to the speaker. Reduced audio quality predictably degrades ASR accuracy. Also, many presentations cover topics with domain specific terms that are not included in vocabularies of general purpose ASR systems.[2] Such domain specific terms can be important search queries.

This paper examines the relative effectiveness of slide text recovered with OCR and spoken text recovered using ASR for indexing lecture videos. We assemble a corpus of presentation videos with automatic and manual transcripts of the spoken text. For these videos, we also have presentation files to generate a manual transcript of the slide text. The TalkMiner system is used to obtain an automatic slide text transcript. This data set allows us to quantify the accuracy of automatic text recovery from both slides and speech within the corpus. Analysis reveals that ASR and OCR errors exhibit different characteristics. Next, we conduct several controlled video retrieval experiments. The results show that slide text enables higher precision video retrieval than spoken text. This reflects both the distinct error profiles of ASR and OCR, as well as basic usage differences between spoken and slide text in presentations. Retrieval combining automatically extracted slide and spoken text achieves improved performance over both individual modalities.

2. RELATED WORK

Early work on video retrieval using spoken text indexed news broadcasts using ASR.[3] News video has been a steady focus of related work, in part because it has greater topic structure than generic video. It also is more likely to contain machine readable text from graphics and tickers common to that genre. Throughout the TRECVID evaluations,[4] news and archival video has been indexed using ASR, at times in conjunction with machine translation. In some cases, the resulting transcripts exhibited high word error rates. Interactive retrieval using visual features alone in a sufficiently powerful interface achieved performance comparable to traditional retrieval using ASR.[5] The Informedia project also applied OCR to video retrieval.[6] Compared to presentation video, the graphics with text used in news broadcasts typically occupy a smaller portion of the frame.

Multimedia retrieval research has examined projector-based slide capture systems that produce a stream of slide frames rather than presentation video (i.e. RGB capture). The resulting stream is lower in complexity and can provide high resolution slide images. In this context, Vinciarelli and Odobez[7] developed techniques to improve OCR's retrieval power, without using ASR. Along similar lines, Jones and Edens[8] described methods for aligning the slides from a presentation file with an audio transcript using a search index constructed from ASR. Slide text is used to query the search index for alignment. They extend this work using a corpus of meeting videos.[9] ASR is used to create a topic segmentation of the meeting, and the slide title queries are used in experiments for exploration of the corpus. Consistently, the accuracy of ASR directly impacts the downstream retrieval performance.

Park *et al.*[2] combined external written sources of specialized vocabulary for improving ASR for lecture indexing. This work showed good results in combination with speaker adaptation and other enhancements. They report that high ASR word error rates can be tolerable, but that video-specific keywords need to be recognized accurately for effective retrieval. Similar experiments were reported elsewhere.[10–12] Swaminathan *et al.*[11] use slide text (from a presentation file) to help correct mistakenly recognized terms in ASR transcripts. They report some reductions in these errors. Other image analysis methods match slides from a presentation file with video frames in which they appear.[13,14]

Most related work focuses on either spoken document retrieval or generic video retrieval. Despite years of steady progress on performance, accuracy continues to pose challenges to the incorporation of ASR in multimedia retrieval systems. While ASR and closed caption (CC) transcripts still generally outperform indexing by content-based visual analysis, slide text recovered by OCR is valuable for indexing presentation videos. We focus on presentation video as a unique genre in which automatically recovered spoken text and slide text can both be individually exploited and combined to improve retrieval.

In the reminder of this paper, we first assess the accuracy of automatically recovered spoken text using ASR and the accuracy of automatically recovered slide text using automatic slide detection and OCR. Secondly, we compare the characteristics of errors in these two modalities. Next, we conduct experiments that examine the impact of transcription errors on video retrieval. Finally, we combine these modalities for lecture video retrieval.

3. DATA PROCESSING

This Section describes the set of videos used in our experiments.

3.1 Data set

Presentation videos from the 2011 Google I/O conference are available on YouTube. These videos comprise an ideal data set due to the availability of manually created closed caption transcripts. ASR transcripts are also provided for these videos by YouTube. These resources represent both ground truth and automatic transcripts for the spoken text. Additionally, PDF, PowerPoint, or HTML presentation slide files are also available for the majority of the presentation videos. These files are processed to create a ground truth slide text transcript. To create the automatic slide text index, we utilize TalkMiner's slide detection system.[1] Text is recovered by applying OCR to the automatically detected slide keyframes.

Our experiments utilize a subset of 74 videos for which all of these data sources are available. Each video is a separate "document" for retrieval, and the total duration of the corpus is more than 64 hours. Search indexes are constructed using the Lucene search library*, and their respective contents are summarized in Table 1.

Table 1 : The information used to build various search indexes

search index	text sources
WEB	metadata available from the video's YouTube web page
OCR	OCR text extracted from automatically detected slide keyframes
SLIDE	text from the presentation slide file
ASR	text from the YouTube ASR transcript
SPOKEN	text from YouTube closed caption spoken transcript
AUTO	ASR transcript, OCR transcript, WEB metadata
MANUAL	text from SLIDE presentation file and manual SPOKEN transcript

After stop word removal, the ASR transcripts contain 22,389 unique terms, while the OCR index contains 58,664 terms. The ground truth slide and spoken text are quite different. The corpus-level intersection ($|\text{SPOKEN} \cap \text{SLIDE}| = 6439$) includes only 49% of SPOKEN terms and 61% of SLIDE terms. The intersection OCR $\cap$ ASR is proportionally smaller still: 14% of OCR terms and 37% of ASR terms. Although conventional web search relies heavily on the metadata in the WEB index, this index contains only 1256 unique terms. The size of the WEB index is less than 6% of the ASR index and less than 3% of the OCR index. For this reason, augmenting metadata-based descriptions with automatically recovered text is critical to support effective lecture video retrieval.

3.2 Transcription Accuracy

Next, we assess the accuracy of automatically extracted slide and spoken text using the manually created ground truth transcripts. There are two types of recognition errors: insertions and deletions. When a slide term is incorrectly recognized by OCR, the resulting erroneous term is inserted in the automatic transcript but is absent from the SLIDE ground truth. Denote the set of OCR insertions for video v by

$$\text{OCR} \setminus \text{SLIDE}(v) = \{t : t \in \text{OCR}(v), t \notin \text{SLIDE}(v)\} \quad .$$

Terms that appear in the ground truth but not in the corresponding automatic transcript are deletions. If a spoken term is not in the ASR system's vocabulary, it appears in the SPOKEN transcript but is missing in the ASR transcript. The set of ASR deletions is

$$\text{SPOKEN} \setminus \text{ASR}(v) = \{t : t \in \text{SPOKEN}(v), t \notin \text{ASR}(v)\} \quad .$$

*http://lucene.apache.org

Insertion errors are far more common than deletions. The OCR transcripts exhibit a higher average insertion rate: 63.5% and 43.7% for OCR and ASR, respectively. Most OCR insertions are due to *character* (i.e. partial word) recognition mistakes, and can occur due to poor quality video capture or small font sizes. On average, more than 88% of the terms in the slide ground truth (SLIDE) appear in the OCR transcripts per video. OCR does not rely on a predetermined vocabulary, so it accommodates specialized terminology. On the other hand, ASR insertion errors are valid (English) words from the ASR system's vocabulary. These *word* recognition mistakes commonly result from phonetic or out of vocabulary mismatch[†]. The average per-video deletion rate of 20.6% for ASR is also higher than for OCR (11.6%).

Table 2 : The average, normalized per-video overlap of automatic (ASR, OCR) transcripts with ground truth (SLIDE, SPOKEN, MANUAL) transcripts and an English language dictionary (EN DICT)

	SPOKEN	SLIDE	MANUAL	EN DICT
OCR	0.324	0.365	0.429	0.489
ASR	0.563	0.146	0.570	0.906

Table 2 shows the average normalized overlap of the OCR transcripts with the SPOKEN and SLIDE ground truth, the combined MANUAL ground truth, and an English language dictionary for each video. In the OCR row, the normalization is relative to the size of each video's OCR transcript, and the average is across all videos. For example, 32% of the OCR text appears in the SPOKEN index. The average overlap with the SLIDE transcripts is 36.5%, but the overlap with the dictionary is only 48.9% due to the large proportion of character-level errors. The overlap with the SPOKEN ground truth is almost as high as with the SLIDE ground truth, indicating that many of the correctly recognized slide terms are also spoken. Finally, the higher overlap between the OCR and the MANUAL transcripts compared to the OCR and SLIDE information indicates that the OCR includes SPOKEN terms absent in the SLIDE index. We attribute this to both OCR errors that produce valid words, and also text recognized in embedded images in the slides that is spoken but not included in the SLIDE transcripts.

The bottom row (ASR) of Table 2 shows analogous results for spoken text. The ASR transcripts' overlaps are higher with the SPOKEN and MANUAL ground truth text than the OCR transcripts. The overlap between the ASR and the SLIDE ground truth is only 14.6%; many terms in the ASR transcript are not in the slides. The great majority of the ASR transcripts' terms, 90.6% on average, appear in the English language dictionary, as ASR mistakes occur largely at the word-level. Restricting analysis to insertion errors, on average 29.9% of OCR insertions ($\text{OCR} \setminus \text{SLIDE}(v)$) are valid dictionary words, while 80.8% of ASR insertions ($\text{ASR} \setminus \text{SPOKEN}(v)$) are valid dictionary words.

Word-level and character-level recognition errors have distinct implications for video retrieval. Because many OCR insertions are not dictionary words, they are unlikely to appear in search queries or in the OCR transcripts of other videos. By contrast, most ASR errors are dictionary words and are more likely to appear in both queries and the ASR transcripts of other videos.

Table 3 quantifies this difference. We assume that SLIDE or SPOKEN ground truth terms are reasonable potential queries for the corresponding videos. The upper table shows the average proportion of insertion errors that appear in the manually created transcript for at least one other video in the corpus. On average, 29% of the erroneously inserted terms in a specific video's OCR transcript are reasonable queries for *some other* video. If those queries are issued, the video with the insertion will be included erroneously among the results. For spoken text, the same rate jumps to 55% of the ASR insertion errors. Thus, while the OCR insertion rate is higher than ASR, a much larger proportion of ASR insertions appear in the manual SPOKEN transcript for at least one other video. This difference between slide and spoken text recognition errors persists for search indexes created from the automatic transcripts. In this case, 51% of OCR insertions appear in the OCR transcript of some other video. For ASR, the overlap jumps to 83%. At retrieval time, the word-level errors of ASR can be expected to degrade precision and overall performance more than the character-level errors of OCR.

[†]Throughout, we use "dictionary word" to indicate a valid word found in the English dictionary. The ASR vocabulary is the pre-defined set of words that the ASR system matches phonetically to the presentation audio stream.

Table 3 : The average, normalized per-video overlap of automatic ASR and OCR insertion errors with four corpus-level sets of terms

	$\cup_{d \neq v}$ SPOKEN(d)	$\cup_{d \neq v}$ SLIDE(d)
OCR\SLIDE(v)	0.277	0.294
ASR\SPOKEN(v)	0.552	0.403

	$\cup_{d \neq v}$ ASR(d)	$\cup_{d \neq v}$ OCR(d)
OCR\SLIDE(v)	0.309	0.512
ASR\SPOKEN(v)	0.839	0.537

4. RETRIEVAL EXPERIMENTS

This section describes retrieval experiments using search indexes constructed from the text sources in Table 1. We assemble a set of synthetic queries from the manually created transcripts. To select queries that are both reliably descriptive and unbiased towards slide or spoken text, we collect the set of text terms that co-occur in each video's SLIDE and SPOKEN transcripts. Combining these terms over the corpus produces the set of 5455 total queries that we use in all experiments. We assess performance using the standard measures of precision, recall, and the F1-score.[15] These are calculated for the set of videos retrieved by the automatic indexes (OCR, ASR, WEB, AUTO) relative to the (relevant) videos returned by the ground truth indexes (SLIDE, SPOKEN, MANUAL) for each query. We report the average of each performance measure over the entire query set.

Table 4 : Slide retrieval results averaged over various ground truth query sets: Results are grouped by headings indicating the search index whose retrieved results are relevant "REL = <ground truth index>". Precision, recall, and F1 score are computed relative to the relevant set of videos and averaged over all queries.

	REL = SLIDE			REL = SPOKEN			REL = MANUAL		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
AUTO	0.42	0.96	0.59	0.66	0.86	**0.75**	0.71	0.86	**0.78**
ASR	0.24	0.59	0.34	0.57	0.64	0.61	0.58	0.60	0.59
OCR	0.70	0.91	**0.79**	0.69	0.50	0.58	0.79	0.53	0.63
WEB	0.09	0.03	0.05	0.14	0.02	0.04	0.15	0.02	0.04
MANUAL	0.54	1.00	0.70	0.92	1.00	0.96	1.00	1.00	1.00
SPOKEN	0.50	0.85	0.63	1.00	1.00	1.00	1.00	0.92	0.96
SLIDE	1.00	1.00	1.00	0.85	0.50	0.63	1.00	0.54	0.70

Table 4 shows all results organized under headings indicating the corresponding ground truth search index used to compute the performance metrics. Boldface entries denote the best performance among the *automatic* search indexes in each condition. The leftmost columns, under "REL = SLIDE", show performance using the videos retrieved by the SLIDE index as relevant results. These results assess the performance of each index for retrieving slide information. As expected, OCR outperforms ASR. The results for the combined automatically recovered text index, AUTO, demonstrate that including ASR degrades precision markedly while only slightly improving recall relative to OCR. Notably, the precision of *ground truth* spoken text (SPOKEN) is only 0.5, while the recall is 0.85. In the "REL=SPOKEN" columns, relevant videos are retrieved by the SPOKEN index. In this condition, the indexes retrieve videos according to spoken information. Here, ASR performs slightly better than OCR and slightly worse than SLIDE. Integrating the OCR information in the AUTO index noticeably improves both precision and recall compared to ASR alone. Here, the precision of SLIDE is 0.85, but the recall is 0.5. The ground truth results demonstrate basic differences in the usage and distinctiveness of slide text and spoken text that are independent of the accuracy of automatic text recovery. Observe that incorporating SPOKEN text lowers the precision of MANUAL when seeking slide information (1-0.54=0.46) much more than adding SLIDE text when seeking spoken information (1-0.92=0.08).

As expected, videos retrieved by the ASR index are consistent with the SPOKEN results. Likewise, videos retrieved by the OCR index are consistent with the SLIDE results. However, OCR performs almost as well as ASR when retrieving spoken information ("REL=SPOKEN"), but ASR performs far worse when retrieving slide information ("REL=SLIDE"). The combined use of OCR and ASR information in AUTO improves retrieval of spoken information, but lowers precision dramatically when retrieving slide information.

We assess performance using the videos retrieved by the combined MANUAL index, in the rightmost columns. The recall values for the SPOKEN and SLIDE indexes show that slide text and spoken text retrieve different videos. The spoken text dominates by volume, as seen in the recall gap (0.92-0.54=0.38). However, this gap is decidedly lower between ASR and OCR (0.60-0.53=0.07). The OCR index shows better F1 performance than ASR, due to a large margin in precision (0.79-0.58=0.21). The superior reliability of automatically recovered slide text is evident in the negligible drop in recall between SLIDE and OCR (0.01), compared to the much greater gap between SPOKEN and ASR (0.92-0.53=0.39). Also, the precision of OCR is uniformly superior to ASR and AUTO. Note that in all conditions, the sparsity of the WEB index produces poor performance. Presentation video retrieval clearly benefits substantially from the use of automatically recovered slide and spoken information.

These results reflect first the content differences between slide and spoken text. This is apparent in the F1 score of 0.63 when using SLIDE to retrieve the SPOKEN results or vice-versa. Also, there is a simple volume ("document length") difference. Spoken transcripts are generally longer since speech is improvised while slides are deliberately authored. The average transcript lengths are 6217 and 3929 terms for ASR and OCR, respectively. Similarly, the lengths for SPOKEN and SLIDE are 6366 and 762 terms, respectively. The average transcript length for the WEB index is only 60 terms.

The differing costs of word-level and character-level recognition errors are also evident. ASR shows higher recall when retrieving spoken information ("REL=SPOKEN","REL=MANUAL"), at the cost of degraded precision. Despite containing substantially less text, the OCR index achieves competitive recall and outperforms ASR in terms of F1 score for general lecture video retrieval ("REL=MANUAL"). OCR exhibits consistently high precision in all conditions.

5. CONCLUSION

The first conclusion is that the slide text and spoken text are not the same. Comparison of the ground truth and automatic transcripts reveal substantial differences in the content and volume of slide and spoken text. The overlap is limited even when controlling for recognition errors with manual transcripts. Issuing term queries that are common to the SLIDE and SPOKEN ground truth retrieve different videos among the results using both manual and automatic text search indexes.

Secondly, both manually and automatically extracted slide text exhibit greater retrieval precision when compared to manually and automatically transcribed spoken text. We attribute this result to two causes. First, the usage of terms in slides is the product of a deliberate authoring process, while speech is often partially improvised. Less descriptive terms are more common in speech, and in turn more commonly shared with other videos' spoken transcripts. This imprecision limits the discriminative power of spoken text for video retrieval. The second factor is the differing recognition error profiles of ASR and OCR. Errors are more frequent in OCR, but occur at the character level producing non-dictionary terms in the transcripts. These errors do not degrade text-based retrieval, since they do not appear as queries. Errors in ASR occur at the word level due to phonetic and out of vocabulary mismatch. The resulting inserted terms tend to be dictionary words that appear in both other video transcripts and search queries.

ASR word-level errors directly impact retrieval performance. The results above show that OCR (F1=0.58) can be almost as effective as ASR (F1=0.61) for video retrieval based on spoken information ("REL=SPOKEN"). At the same time, OCR is vastly superior to ASR for retrieval of slide information ("REL=SLIDE"). OCR also outperforms ASR in the general case despite containing less text than the ASR index ("REL=MANUAL").

Combining these two modalities represents an opportunity for improving presentation video retrieval. A combined index that incorporates slide text improves retrieval precision over using spoken text alone. Similarly, incorporating spoken text can improve the retrieval recall over using slide text alone. Building on these insights, we intend to design an indexing and retrieval strategy to leverage the relative strengths of both automatically recovered slide and spoken text.

REFERENCES

[1] Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H., and Rowe, L. A., "Talkminer: a lecture webcast search engine," in [*Proceedings of the international conference on Multimedia*], *MM '10*, 241–250, ACM, New York, NY, USA (2010).

[2] Park., A., Hazen, T. J., and Glass, J. R., "Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling," in [*Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*], **1**, 497–500 (2005).

[3] Brown, M. G., Foote, J. T., Jones, G. J. F., Sparck Jones, K., and Young, S. J., "Automatic content-based retrieval of broadcast news," in [*Proceedings of the third ACM international conference on Multimedia*], *MULTIMEDIA '95*, 35–43, ACM, New York, NY, USA (1995).

[4] Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A. F., and Quéenot, G., "Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in [*Proceedings of TRECVID 2011*], (2011).

[5] Adcock, J., Cooper, M., and Pickens, J., "Experiments in interactive video search by addition and subtraction," in [*Proceedings of the 2008 international conference on Content-based image and video retrieval*], *CIVR '08*, 465–474, ACM, New York, NY, USA (2008).

[6] Hauptmann, A., Jin, R., and T.D.Ng, "Video retrieval using speech and image information," in [*SPIE Storage and Retrieval for Multimedia Databases 2003, EI'03 Electronic Imaging, Santa Clara, CA, January 20-24th, 2003.*], **5021** (2003).

[7] Vinciarelli, A. and Odobez, J.-M., "Application of information retrieval technologies to presentation slides," *IEEE Trans. on Multimedia* **8**(5), 981–995 (2006).

[8] Jones, G. J. F. and Edens, R. J., "Automated alignment and annotation of audio-visual presentations," in [*Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*], *ECDL '02*, 276–291, Springer-Verlag, London, UK, UK (2002).

[9] Jones, G. J. F., Eskevich, M., and Gyarmati, A., "Towards methods for efficient access to spoken content in the ami corpus," in [*Proceedings of the 2010 international workshop on Searching spontaneous conversational speech*], *SSCS '10*, 27–32, ACM, New York, NY, USA (2010).

[10] Kawahara, T., Nemoto, Y., and Akita, Y., "Automatic lecture transcription by exploiting presentation slide information for language model adaptation," in [*IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*], 4929–4932 (2008).

[11] Swaminathan, R., Thompson, M. E., Fong, S., Efrat, A., Amir, A., and Barnard, K., "Improving and aligning speech with presentation slides," in [*Proceedings of the 2010 20th International Conference on Pattern Recognition*], *ICPR '10*, 3280–3283, IEEE Computer Society, Washington, DC, USA (2010).

[12] Matton, M. and Braeckman, K., "Speech recognition tools in a media retrieval system," in [*Proceedings of the 2011 ACM international workshop on Automated media analysis and production for novel TV services*], *AIEMPro '11*, 19–24, ACM, New York, NY, USA (2011).

[13] Erol, B., Hull, J. J., and Lee, D.-S., "Linking multimedia presentations with their symbolic source documents: algorithm and applications," in [*Proceedings of the eleventh ACM international conference on Multimedia*], *MULTIMEDIA '03*, 498–507, ACM, New York, NY, USA (2003).

[14] Tung, Q., Swaminathan, R., Efrat, A., and Barnard, K., "Expanding the point: automatic enlargement of presentation video elements," in [*Proceedings of the 19th ACM international conference on Multimedia*], *MM '11*, 961–964, ACM, New York, NY, USA (2011).

[15] Manning, C. D., Raghavan, P., and Schütze, H., [*Introduction to Information Retrieval*], Cambridge University Press (2008).