Generating Automatic Keywords for Conversational Speech ASR Transcripts

Hohyon Ryu
Twitter, Inc.
San Francisco, CA 94103
hohyonryu@twitter.com

Matthew Lease
School of Information
University of Texas at Austin
ml@ischool.utexas.edu

ABSTRACT

While a plethora of *conversational speech* has been recorded and archived for over a century, it has not been easily accessible due to many technical challenges vs. text and *rehearsed speech* to be addressed before conversational archives can be effectively searched and used. In this paper, we describe two language modeling methods for automatically assigning keywords to automatic speech recognition (ASR) transcripts, to benefit search and browsing of conversational speech archives. Experiments performed with the English CLEF CL-SR MALACH collection of oral history interviews. In comparison to a prior baseline generating 20 keywords per conversation segment, we use 1/20th the training data yet improve Recall@20 in matching manual keywords. However, while indexing of manual keywords yields improved search accuracy, indexing automatic keywords (ours or the baseline) fails to improve search accuracy, evidencing the need for additional research.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*

General Terms

Algorithms, Experimentation, Performance

Keywords

Spontaneous Conversational Speech, Language Modeling, Nearest Neighbor Classifier

1. INTRODUCTION

While *spoken document retrieval* was claimed to be a solved problem [3] in TREC over a decade ago, the data considered was only clean *prepared speech*, aka "rehearsed" or "read" speech, such as broadcast news and political speeches. In contrast, *spontaneous conversational speech* (SCS) (e.g., phone calls or voicemail, meetings, classroom discussion, talk shows, interviews, cocktail parties, etc.), which has been widely collected and archived for over a century, has mostly remained in its raw form posing many challenges for effective information retrieval (IR) today [11, 16, 10, 13].

With SCS, automatic speech recognition (ASR) performs much worse than with prepared speech due to many factors: wider speaker variation, non-native speakers who may "code-switch" between languages, emotional speech, noisy environments, and lower quality microphones [5, 4]. Speech may also use specialized vocabulary not part of common discourse and ASR vocabularies. In addition, speech exhibits cognitive processing effects and speaking errors, including mumbles, partial-words, filler words, repetitions, and corrections. The following ASR transcript provides an example of ASR quality in the MALACH oral history collection [11] (excerpted from VHF34774-159541.027, see Section 3.1):

> do you recall the first time you were beaten yes forty forty five what was that did you what it was not the the the the the cement or do you it was very hard for me was at home uh one of the people live like that you know and that was not far from the city where you couldn't you was punished at all is that the that the from the i could go back to the were couldn't buy any friends as a uhhuh polish and what did you so when you get the uhhuh lucky superficial watches who did such a you and if you could have uhhuh and what and that was what was your what and that was all in and polish jews remember his name in the face of all slammed hide the fact that you...

When relevance judgments were being collected for the English MALACH IR test collection, ASR transcripts were not yet available. Instead, judging relied on manual summaries and keywords, with occasional reference to the audio. Metadata for this same conversation segment is show below:

> Auschwitz II-Birkenau (Poland : Death Camp) | Poland 1944 | kapos, Jewish | kapos, Polish | brutal treatment in the camps | beatings

For the search topic "Birkenau daily life", "Birkenau" never appears in the ASR text but does appear in the manual keywords, and this segment was indeed judged as relevant to this search topic. How is a user or search system to recognize this segment's relevance without labor-intensive manual curation to produce such keywords? To address this, prior work has investigated automatic keyword generation [16,

ACM SigIR '2013 Dublin, Ireland

12]. We further explore this idea, building on our prior inferring a text's place and time from implicit lexical cues rather than detecting explicit places or dates [8, 17, 14].

From 2005-2007, the Cross Language Evaluation Forum (CLEF) held a Cross-Language Speech Retrieval (CL-SR) Track [16, 10, 13]. The CL-SR track used part of the Survivors of the Shoah Visual History Foundation oral history archive, and the retrieval tasks were conducted on the ASR text, manual summary, manual keywords, and automatic keywords. We study the CL-SR'07 test collection here.

In [2] and [15], a machine learning approach was proposed for keyword extraction. [1] applied a keyword extraction algorithm designed for written text and showed ASR quality is crucial to keyword extraction performance. [6] extracted keywords using supervised machine learning and linguistic knowledge. [7] further refined [1]'s method taking into account the semantic meanings of the keywords. [9] extracted keywords from a conversational meeting corpus using supervised machine learning approach and bigram expansion.

The most relevant prior work [16, 12] to ours inferred keywords for the same MALACH collection. [12] used the previous segment to provide additional context for the next segment. However, whereas their data-intensive approached utilized 168,584 segments of private training set, we attempt this task using only the roughly 8,000 conversation segments publicly available in the CL-SR'07 collection.

2. METHODOLOGY

Our task is to automatically select the best keywords, from a pre-defined set, to assign to each conversational segment. We describe two approaches to this task below. In both, a language model (LM) estimates a probability of document d being relevant to query q, or $p(d|q)$. Using Bayes Rule, we derive the usual query-likelihood IR formulation as:

$$p(d|q) = \frac{p(q|d)p(d)}{p(q)}$$

$p(q)$ can be ignored since it is constant across documents. $p(d)$ is the document prior. $p(q|d)$ is query likelihood, i.e. the probability of generating query q for document d.

2.1 Pseudo-Document (PD) Language Model

In our first approach, we construct a pseudo-document for each keyword by aggregating all the segments to which the given keyword has been manually assigned. In this way, we create a collection of psuedo-documents, one per keyword, which can be searched. Each conversation segment represents a query, and the top-K ranked pseudo-documents correspond to the most likely K keywords that should be assigned, where K is a parameter. The LM is redefined as $p(k|s)$ for keyword k and conversation segment s via Bayes:

$$p(k|s) = \frac{p(s|k)p(k)}{p(s)}$$

where denominator $p(s)$ is again constant and can be ignored, $p(k)$ defines a keyword prior, and $p(s|k)$ is the likelihood of the segment given a particular keyword.

We investigate expanding the query segment with similar segments to provide contextual information (CI). For segment s with word vector $\vec{w}$, neighbor segment s' is added to s with weight inversely-proportional to its distance from s.

$$\delta = |i_{s'} - i_s|, w'_j = f(d; \mu, \sigma^2) \times tf_{w'}$$

where δ is the distance between s and s', i is the sequential segment index, $tf_{w'}$ is the term frequency of a word w' in the segment s', and f is a Gaussian probability density function:

$$f(x;\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/(2\sigma^2)}$$

As in prior work [16, 12], we also try assigning keywords with two separate models (TM): a temporal-geolocation model and a more general concept model. Each model is trained by partitioning the manual keywords into these two categories. For example, 'Germany 1943' vs. 'literature and writing'.

2.2 Segment Language Model (kNN)

Prior work [16, 12] used k-Nearest neighbor (kNN) approach to find similar segments and their keywords. We explore a similar LM approach here, using the query segment to find similar segments (from other interviews).

$$p(s'|s) = \frac{p(s|s')p(s')}{p(s)}$$

We then aggregate manual keywords assigned to those similar segments in order to select keywords to assign to the query segment. Given the top ranked k most similar segments, each manual keywords assigned to the ith retrieved segment s_i is receives weight $w_{kwd} = k - i + 1$. The n keywords with highest aggregated weight are assigned to s.

As with the earlier psuedo-document approach, the query segment can be expanded with similar segments prior to matching, akin to traditional IR psuedo-relevance feedback.

3. EXPERIMENTS

Search to infer keywords is performed using Galago[1]. With the two model (TM) approach, separate searches are used for temporal-geolocation vs. general concept keywords. 16 concept keywords and 4 temporal-geolocation keywords are assigned to each segment. We evaluate in two ways: automatic vs. manual keywords, and change in CLEF CL-SR search accuracy when we add keyword indexing.

3.1 Test Collection

Field	Description
DOCNO	the interview id + the segment id
INTERVIEWDATA	the names of interviewees
NAME	the names of people mentioned in the segment
ASRTEXT2003 ASRTEXT2004 ASRTEXT2006A ASRTEXT2006B (ASR)	4 versions of ASR texts
SUMMARY (SUM)	manual summary
MANUALKEYWORD	manual keyword
AUTOKEYWORD2004A1 AUTOKEYWORD2004A2	2 versions of automatic keywords

Table 1: CLEF CL-SR Interview Segment Fields.

The CLEF 2007 Cross-Language Speech Retrieval collection (CL-SR) [13] is used in this experiment. CL-SR consists of 272 interviews with Holocaust survivors, witnesses, and rescuers (589 hours of speech) [16]. These interviews are divided into 8,104 segments, including four versions of ASR

[1] www.galagosearch.org

transcripts (we use the best only, ASRTEXT2006B), manual keywords, and two versions of automatic keywords. Table 1 shows the metadata fields for each segment. We exclude two interviews (15 segments) which are missing ASR texts. Short, blank, or corrupted ASRs are also filtered out, leaving 7902 segments.

We observe that 5.6 manual keywords are assigned on average, using 3605 unique keywords. Of the two sets of automatic keywords, we adopt AUTOKEYWORD2004A2 as baseline since it better matches manual keywords. Quality of baseline keywords are shown in Table 2. As the baseline methods assign 20 keywords per segment, we report Recall at 20 keywords (R@20) for comparable evaluation.

Keywords	**Precision**	**Recall**	**F-Score**
AUTOKEYWORD2004A1	0.076	0.289	0.116
AUTOKEYWORD2004A2	0.090	0.326	0.136

Table 2: CL-SR baseline automatic keyword quality.

3.2 Methods

Pseudo-documents. Keywords are assigned to the segments using 10-fold cross-validation. We retrieve the 20 most similar keyword pseudo-documents for each query segment s. The query segment s is simply the segment's ASR transcript. For each keyword pseudo-document k, we not only concatenate the ASR transcripts for the segments to which it is assigned, but we also follow prior work [16, 12] in "cheating" by including manual summaries as well. This allows fair comparison but will ultimately be abandoned in future work. Nevertheless, the task remains quite difficult.

Similar segments. Using Galago, similar segments s' are retrieved for query segment s. Segments s' in the Galago search index includes both ASR text and the manual summary, whereas the query segment includes ASR text only. Experiments vary the number of similar segments used.

3.3 Searching MALACH

Whereas our first evaluation method assesses the system's ability to match manual keywords, our second evaluation measures the benefit of automatic keywords for improving search accuracy. We use 105 CL-SR topics, where queries include all topic fields: title, description, and narrative (e.g., see Table 3). Gold relevance judgments are binary [13]. Stopwords (ST) are removed based on the Indri[2] stop word list, augmented to exclude conversational filled pauses and backchannels such as: "um", "yeah", "uhhuh", and "wow".

Search uses ElasticSearch[3] 0.19.9, based on Lucene[4] 3.5. Retrieval performance, measured by Mean Average Precision (MAP), is compared with keywords we inferred is compared to use of the CL-SR manual or automatic keywords.

4. RESULTS

4.1 Matching Manual Keywords

Table 4 shows results. Critically, note that that BASELINE used 168,584 segments and kNN used about 7,000 segments for training [16, 12], whereas we use only a few thousand segments. We see the kNN approach far outperforms the pseudo-document approach. Recall ST denotes stopwords filtering, CI is context information, TM is the two model approach, and k# denotes the value of parameter k of kNN (Section 2.2). Adding context information led to detrimental query drift and significantly decreased the keyword matching performance. TM had no significant impact.

Topic	Varian Fry
Description	The story of Varian Fry and the Emergency Rescue Committee who saved thousands in Marseille
Narration	Varian Fry, a young American journalist, created an underground operation that smuggled more than 2,000 refugees (including Marc Chagall, Max Ernst, and Andre Breton) out of Vichy France in 1940-1941. The relevant material should contain information about this operation. Any first-hand information of people who have been rescued by Fry is highly relevant

Table 3: A example of the topics.

Experiment	**R@20**
BASELINE	0.334
PD	0.047
PD(ST)	0.084
PD(ST+CI)	0.029
PD(ST+TM)	0.085
kNN(k10)	0.235
kNN(k10+ST)	0.276
kNN(k200+ST)	**0.369**
kNN(k200+ST+CI)	0.32

Table 4: Keyword matching performance measured for Recall at 20. Only stopword filtered kNN approach with a large k outperformed the baseline.

Figure 1 shows the impact of k of kNN. kNN with stopword filtering (ST) outperforms the baseline at 40.

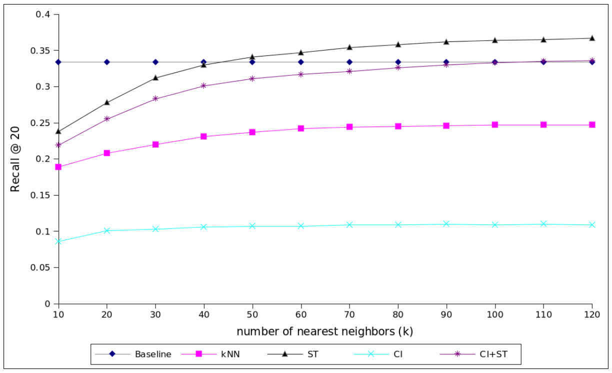

Figure 1: The IR Test Performance of the MANUALKEYWORD and the baselines (AUTOKEYWORDS2004A1 and AUTOKEYWORDS2004A2).

4.2 Using Automatic Keywords in Search

Figure 2 shows the impact of adding keywords for IR performance. We index SUMMARY, ASRTEXT2006B and an additional keyword field that is varied: MANUALKEYWORD, AUTOKEYWORD2004A2 (baseline), and finally

[2] `www.lemurproject.org/indri`

[3] `www.elasticsearch.org`

[4] `lucene.apache.org`

our kNN(k200+ST) generated keywords. We vary the number of keywords added from kNN(k200+ST) from 5 to 20.

Manual keywords improve the IR performance significantly while the automatic keywords have almost no impact. The baseline AUTOKEYWORD2004A2 has zero or negative impact in comparison to the NO KEYWORDS condition. The impact of kNN(k200+ST) on IR performance is negligible.

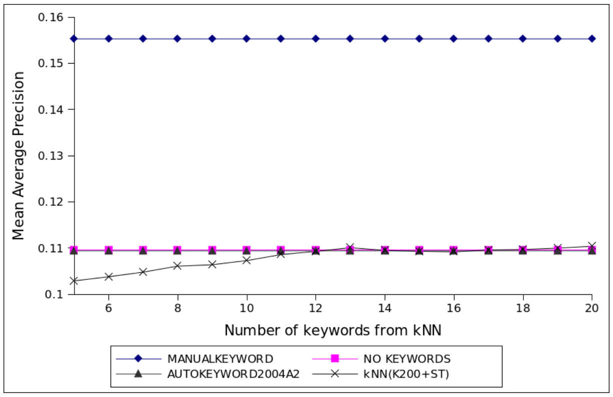

Figure 2: The IR Test Performance of SUMMARY and ASRTEXT2006B for CLEF 2007 topics, descriptions, and narrations when MANUALKEYWORD, AUTOKEYWORDS2004A2, kNN(k200+ST), and no keywords are added.

5. DISCUSSION

We compared two language modeling approaches to improve SCS retrieval: generating a psuedo-document for each keyword, and assigning keywords from similar segments. Inferred keywords were also evaluated in terms of change in IR performance using CLEF 2007 CL-SR track IR tasks. For matching manual keywords, we were able to improve Recall@20 despite training on roughly 20x fewer segments than used in prior work. Nevertheless, neither our generated keywords, nor the baseline automatic keywords, led to improved IR accuracy when indexed.

Presently we see that in the majority of the cases, automatic keywords introduce more incorrect than correct keywords. Out of 20 keywords, only about 10% of the automatic keywords are correct, leaving 90% of keywords to confuse the search engine. This is familiar to traditional NLP approaches in that while latent representations offer opportunities to enrich observed terms, errors inferring latent structures can cause more harm than good. Thus, instead of measuring recall, we should really be focusing on improving the precision of keyword extraction to benefit IR accuracy.

As this is a work-in-progress, we have a variety of ideas for refining the modeling approaches from here, and we are also looking into whether the additional training segments used in prior work [16, 12] might be obtainable.

Acknowledgments

This research was supported in part by DARPA Award N66001-12-1-4256, IMLS grant RE-04-13-0042-13, and a Temple Fellowship. Any opinions, findings, and conclusions or recommendations expressed by the authors do not express the views of any of the supporting funding agencies.

6. REFERENCES

[1] A. Désilets, B. D. Bruijn, and J. Martin. Extracting Keyphrases from Spoken Audio Documents. In *Information Retrieval Techniques for Speech Applications*, volume 2273 of *Springer LNCS*, pages 36–50. 2002.

[2] E. Frank, G. Paynter, and I. Witten. Domain-specific keyphrase extraction. In *6th International Joint Conference on Artificial Intelligence*, 1999.

[3] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The TREC spoken document retrieval track: A success story. In *TREC 8*, pages 16–19, 2000.

[4] W. Ghai and M. G. College. Literature Review on Automatic Speech Recognition. *International Journal of Computer Applications*, 41(8):42–50, 2012.

[5] B. Gold, N. Morgan, and D. Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Wiley, 2011.

[6] A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Empirical Methods in Natural Language Processing*, 2003.

[7] D. Inkpen and A. Désilets. Extracting semantically-coherent keyphrases from speech. *Canadian Acoustics*, pages 2–3, 2004.

[8] A. Kumar, M. Lease, and J. Baldridge. Supervised language modeling for temporal resolution of texts. In *ACM Conference on Information and Knowledge Management (CIKM)*, pages 2069–2072, 2011.

[9] F. Liu and Y. Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. *Spoken Language Technology Workshop*, pages 181–184, 2008.

[10] D. Oard, J. Wang, G. Jones, and R. White. Overview of the CLEF-2006 cross-language speech retrieval track. In *CLEF 2006*, pages 744–758, 2007.

[11] D. W. Oard, D. Doermann, G. C. Murray, J. Wang, M. Franz, and S. Gustman. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech Categories and Subject Descriptors. In *27th ACM-SIGIR*, pages 41–48, 2004.

[12] J. S. Olsson and D. W. D. Oard. Improving text classification for oral history archives with temporal domain knowledge. In *Proceedings of the SIGIR*, 2007.

[13] P. Pecina, P. Hoffmannova, and G. Jones. Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 1–14. 2008.

[14] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised Text-based Geolocation Using Language Models on an Adaptive Grid. *Proc. EMNLP-CoNLL*, pages 1500–1510, 2012.

[15] P. D. Turney. Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, 2(3):303–336, 2000.

[16] R. W. White, D. W. Oard, G. J. Jones, D. Soergel, and X. Huang. Overview of the clef-2005 cross-language speech retrieval track. pages 744–759. Springer, 2006.

[17] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *Proceedings of ACL*, pages 955–964, 2011.