INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Incorporating Acoustic Features for Spontaneous Speech driven Content Retrieval

Hiroto Tasaki[1], *Tomoyosi Akiba*[1]

[1]Department of Computer Science and Engineering
Toyohashi University of Technology

tasaki@nlp.cs.tut.ac.jp, akiba@nlp.cs.tut.ac.jp

Abstract

A speech-driven information retrieval system is expected to be useful for gathering information with greater ease. In a conventional system, users have to decide on the contents of their utterance before speaking, which takes quite a long time when their request is complicated. To overcome that problem, it is required for the retrieval system to handle a spontaneously spoken query directly. In this work, we propose an extension technique of spoken content retrieval (SCR) for effectively using spontaneously spoken queries. Acoustic features of meaningful terms in the retrieval may have prominence compared to other terms. Also, those terms will have linguistic specificity. From this assumption, we predict the contribution of terms included in spontaneously spoken queries using acoustic and linguistic features, and incorporate it in the query likelihood model (QLM) which is a probabilistic retrieval model. We verified the effectiveness of the proposed method through experiments. Our proposed method was successful in improving retrieval performance under various conditions.

Index Terms: spoken content retrieval, spontaneously spoken query, acoustic features, linguistic features

1. Introduction

Due to the progress of speech recognition technology and the spread of smart phones in recent years, a voice input interface has become more and more familiar. Also in information retrieval, systems that handle speech input have been developed. They are expected to make it easy for users to gather information. However, current speech-driven retrieval systems simply accept no more than the text input transcribed from its spoken form. Therefore, users have to carefully arrange their query in advance, which take quite a long time when their request is complicated.

On the other hand, speech is one of most natural, easy, and efficient ways for humans to express their own ideas. Therefore, it can be said that the current retrieval systems do not fully make use of such advantages that come from the nature of speech. In order to make use of these advantages, speech-driven retrieval systems are required to accept a spontaneously spoken query instead of a well-arranged counterpart. One of the advantages of such spontaneously spoken queries as input to a retrieval system is that it enables users to easily submit long queries giving systems rich clues for retrieval, although their spontaneous nature means that they are harder to recognize reliably.

In this work, we investigate the speech-driven retrieval system that accept a spontaneously spoken query directly from users. One of the problems accompanied with such systems is in how to distinguish important keywords for retrieval from less useful ones often appearing in spontaneous speeches.

In order to solve this problem, we propose to make use of acoustic features for information retrieval. We define a *contribution for retrieval* for each term, then introduce the prediction model. The estimated contribution is used to enhance a retrieval model to improve retrieval performance. Through our experimental evaluation, we found that incorporation of acoustic features in a spontaneously spoken query successfully improved retrieval performance compared with our baseline.

This paper is organized as follows. Section 2 presents related work. Section 3 explains the conventional SCR method. Section 4 describes our technique to estimate term contribution using acoustic and linguistic features. Section 5 presents experimental results. Finally, Section 5 reports our conclusions and outlines directions for future work.

2. Related work

Researches on spontaneous speech have been actively investigated in the field of spoken language processing thanks to speech corpus where spontaneously spoken utterances are collected, e.g. Switchboard corpus [1] which recorded telephone conversations in American English, Buckeye corpus [2] which recorded utterance in the interview, and the Corpus of Spontaneous Japanese (CSJ) [3] which recorded various type of spontaneous speeches such as academic presentations and simulated dialog.

Spoken Content Retrieval (SCR) is the task of information retrieval targeting speech data. The SCR task was investigated in the MediaEval Benchmark Spoken Web Search Task [4]. In this task, the audio content is spontaneous speech that has been created over the phone in a live setting by low-literate users. In this particular case, most of the audio content is related to farming practices and consists of various languages. The SCR task targeting spontaneously spoken documents in Japanese was investigated in the NTCIR-9 SpokenDoc [5] and the NTCIR-10 SpokenDoc-2 [6]. After that, the SCR task using spontaneously spoken queries in Japanese was investigated in the NTCIR-11 SpokenQuery&Doc [7] and the NTCIR-12 SpokenQuery&Doc-2 [8].

Some researchers have tried to introduce acoustic features in SCR and speech-driven information retrieval tasks. With regard to speech-driven retrieval, Mishra et al. [9] reported that improvements were obtained by using acoustic features in the task of detecting prominent terms in spontaneous spoken queries. With regard to SCR, Racca et al. [10] also made use of acoustic features to construct a classifier of prominent terms in target spoken documents by a support vector machine, where the classified prominences were incorporated into the vector space retrieval model. In spite of these works, the effect of introducing acoustic features into information retrieval has been reported with limited success so far.

http://dx.doi.org/10.21437/Interspeech.2017-893

3. Spoken Content Retrieval

3.1. Conventional SCR method

In a general SCR system, firstly, transcription of spoken documents is acquired by applying ASR. Next, morphological analysis and stop words removal are performed on them, and then they are converted into a set of terms. In the same way, the search query is also converted into a set of terms. In this work, MeCab [11] with UniDic [12] dictionary is used for morphological analysis. Finally, the relevance between the query and each target document is calculated by using a retrieval model, then output the document with high relevance as the retrieval result. We use Query Likelihood Model (QLM) [13] as the retrieval model.

In the query likelihood model, the relevance between document d and query q is expressed as a probability model, where document retrieval is treated as a problem of estimating it. Likelihood $P(q|d)$ of generating a query q from each document d is calculated, then the document having maximum likelihood is selected.

$P(q|d)$ is calculated by Equation 1.

$$P(q|d) \propto \sum_{i \in q} TF(i,q) * \log P(i|d) \tag{1}$$

where i indicates an index of a term in q. The probability $P(i|d)$ of generating a term i from a document d is calculated by Equation 2.

$$P(i|d) = \frac{|d|}{|d|+\mu} P_{ML}(i|d) + \frac{\mu}{|d|+\mu} P_{ML}(i|C) \tag{2}$$

$$P_{ML}(i|d) = \frac{TF(i,d)}{|d|} \tag{3}$$

$$P_{ML}(i|C) = \frac{\sum_{c \in C} TF(i,c)}{\sum_{c \in C} |c|} \tag{4}$$

where $P_{ML}(i|d)$ indicates the maximum likelihood estimate of generating i from d, $P_{ML}(i|C)$ indicates that of generating i from all the document collection. In Equation 2, Dirichlet smoothing [14] is applied to cope with the zero frequency problem, where μ denotes a Dirichlet coefficient. In Dirichlet smoothing, the rate of interpolation is determined by the document length. This means that a short document has a higher necessity for smoothing a long document.

3.2. Extension of QLM using term weight

We extend the original QLM to cope with *contribution for retrieval* of each term in a given query. The method of estimating it is discussed in section 4.

We denote a weight of the *contribution for retrieval* as $W(i,q)$, the value of which is $0 \leq W(i,q) \leq 1$. It is incorporated into the QLM as shown in Equation 5.

$$\begin{aligned} P(q|d) \propto & \alpha \sum_{i \in q} TF(i,q) * \log P(i|d) \\ & + (1-\alpha) \sum_{i \in q} W(i,q) * TF(i,q) * \log P(i|d) \end{aligned} \tag{5}$$

The first and the second term represent the original QLM and the QLM with weight extension, respectively. α represents a parameter that is introduced to balance the two terms. We set it to 0.5 so that both influences are equal. Figure 1 shows the proposed framework of SCR using a spontaneously spoken query.

4. Term weight estimation using acoustic and linguistic features

In order to estimate the weight for terms in the query, we constructed a prediction model by using machine learning with acoustic and linguistic features. This section explains the model construction process.

4.1. Extracted acoustic and linguistic features

We extract the acoustic and linguistic features of each term in the query in constructing the prediction model. In the acoustic features, we investigated two kinds of sets: *plain set* and *emotion set*. In the *plain set*, we extracted the following five types of features using Praat [15].

- average and standard deviation of pitch
- average and standard deviation of intensity
- speech rate

In the *emotion set*, we extracted the features of 384 dimensions used in the task interspeech 2009 emotion challenge [16] for analyzing the emotions of utterances using openSMILE [17]. We also included speech rate in it. Detailed interspeech 2009 emotion challenge features are presented in Table 1 and Table 2.

The extracted features were normalized so that the average and the variance are to be 0 and 1 respectively for each utterance.

In the linguistic features, we extracted the following two types of features.

- surface (convert to one-hot vector)
- Inverse Document Frequency (IDF)

IDF was normalized so that average and the variance are to be 0 and 1 respectively in the entire train query set. We also considered the use of Part Of Speech(POS), but excluded it from the set because good results were not obtained in the preliminary experiment.

4.2. Contribution for retrieval

Contribution for retrieval $C(i,q)$ of each term i in a query q is defined by Equation 6 using the triples: a document collection, a query and a set of relevant documents for it.

$$C(i,q) = \log \frac{P(i|R_q)}{P(i|C)} \tag{6}$$

where $P(i|R_q)$ indicates the occurrence probability of term i in the set of the relevant document set R_q for a query q.

$$P(i|R_q) = \lambda P_{ML}(i|R_q) + (1-\lambda) P_{ML}(i|C) \tag{7}$$

Table 1: *low-level descriptions for interspeech 2009 emotion challenge*

pcm RMSenergy	Root-mean-square frame energy
mfcc	Mel-Frequency cepstral coefficients 1-12
pcm zcr	Zero-crossing rate of time signal (frame-based)
voiceProb	The voicing probability computed from the ACF
F0	The fundamental frequency computed from the Cepstrum
a first-order delta coefficient of the above features	

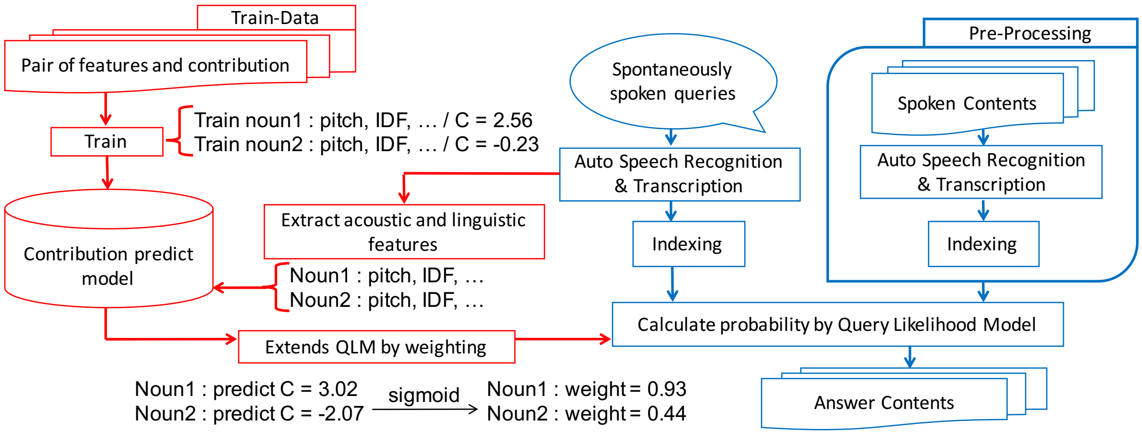
Train-Data
Pair of features and contribution
Train
Train noun1 : pitch, IDF, … / C = 2.56
Train noun2 : pitch, IDF, … / C = -0.23
Contribution predict model
Extract acoustic and linguistic features
Noun1 : pitch, IDF, …
Noun2 : pitch, IDF, …
Extends QLM by weighting
Noun1 : predict C = 3.02
Noun2 : predict C = -2.07
sigmoid
Noun1 : weight= 0.93
Noun2 : weight= 0.44
Spontaneously spoken queries
Auto Speech Recognition & Transcription
Indexing
Pre-Processing
Spoken Contents
Auto Speech Recognition & Transcription
Indexing
Calculate probability by Query Likelihood Model
Answer Contents

Figure 1: *Proposed framework. The blue line shows the structure of the conventional SCR, and the red line shows the proposed extension.*

Table 2: *functionals for interspeech 2009 emotion challenge*

max	The maximum value of the contour
min	The minimum value of the contour
range	= max-min
maxPos	The absolute position of the maximum value (in frames)
minPos	The absolute position of the minimum value (in frames)
amean	The arithmetic mean of the contour
linregc1	The slope (m) of a linear approximation of the contour
linregc2	The offset (t) of a linear approximation of the contour
linregerrQ	The quadratic error computed as the difference in the linear approximation and the actual contour
stddev	The standard deviation of the values in the contour
skewness	The skewness (3rd order moment)
kurtosis	The kurtosis (4th order moment)

$$P_{ML}(i|R) = \frac{\sum_{r \in R_q} TF(i,r)}{\sum_{r \in R_q} |r|} \tag{8}$$

The linear interpolation coefficient λ is set to 0.95. We defined the equation so that the contribution of the term that appeared more frequently in the relevant document set but less in the entire target document set would be higher.

4.3. Estimation of contribution for retrieval

A training sample was constructed for each term by pairing the extracted acoustic and linguistic features with the contribution for retrieval by using the training data. We trained the regression model to predict the contribution for retrieval of a term i in a query q by Support Vector Regression(SVR) [18] with Sequential Minimum Optimizer(SMO) algorithm [19] using data mining software WEKA [20].

The trained regression model is used to predict the contribution of each term in an unseen query in test data. The predicted contributions are transformed to fit into the weights $W(i,q)$ in equation 5 by using sigmoid function as shown in equation 9.

$$W(i,q) = \frac{1}{1 + e^{-C(i,q)}} \tag{9}$$

5. Experiment

5.1. Data

Our experimental evaluation was carried out on the data set provided in the NTCIR-12 SpokenQuery&Doc-2 task [8]. The spoken query set consists of 80 queries of formal-run, which were spontaneously spoken queries including a lot of fillers and hesitations collected from 22 subjects. The target document collection is The Corpus of 1st to 7th Spoken Document Processing Workshop (SDPWS1to7), which consists of 98 lectures with slide-change annotation. A search unit of the collection is called a slide group segment (SGS), which corresponds to a single presentation topic explained by using one or a few presentation slides. Their manual and automatic transcription on both the queries and the documents are provided from the task organizers. Julius [21], a large vocabulary continuous speech recognition system, was used for obtaining the automatic transcriptions. The detail of the dataset is found in [8]. We constructed the retrieval system on the data by using only nouns as indexing terms used in the retrieval model shown in Section 3.

To train our prediction model for *contribution for retrieval*, we constructed the training data from the NTCIR-9 SpokenDoc task [22], whose target documents are 2702 academic lectures in the Corpus of Spontaneous Japanese (CSJ) [3]. We constructed 39 spontaneously spoken queries from the written queries in the testdata by using the method described in [23]. The prediction model was constructed using only nouns in the queries by using their manual transcriptions, since our retrieval model considers only nouns for calculating relevancy.

650 terms were included in the training set, while 2336 and 2408 terms were included in the manual and automatic transcription respectively in the test set.

5.2. Feature extraction

Acoustic features are to be extracted from a speech segment related to the term in question. We tried three kinds of such a speech segment.

term The speech segment just corresponding to the term itself. For automatic transcription of a query, we used the starting and ending timestamps obtained by ASR. For manual transcription of a query, we performed force-alignment between the audio query and its manual transcription to

Table 3: *MAP values obtained by compared retrieval models with and without the proposed extension*

			acoustic features = *plain set*			acoustic features = *emotion set*			
query	document	base-QLM	term	clause	IPU	term	clause	IPU	oracle
manual	manual	0.232	0.243	0.243	0.243	**0.244**	0.234	0.238	0.307
manual	ASR	0.203	0.205	0.205	0.205	**0.217**	0.208	0.204	0.240
ASR	manual	0.192	**0.194**	0.188	0.188	0.193	0.190	0.189	0.246
ASR	ASR	0.182	0.188	0.188	0.188	**0.193**	0.186	0.192	0.217

Table 4: *Varitation in MAP by dividing acoustic features and linguistic features. Column “all” is equal to column 7 in Table 3.*

query	document	all	linguistic	acoustic
manual	manual	**0.244**	0.243	0.231
manual	ASR	**0.217**	0.205	0.204
ASR	manual	**0.193**	0.188	0.190
ASR	ASR	**0.193**	0.188	0.185

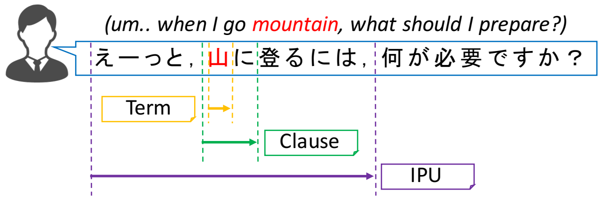
(um.. when I go mountain, what should I prepare?)
えーっと，山に登るには，何が必要ですか？
Term
Clause
IPU

Figure 2: *Method for extracting acoustic features to be attached to nouns (mountain)*

obtain the timestamps.

clause A Japanese sentence consists of a sequence of BUNSETSU, which is defined as one or more content words followed by functional words. We used a BUNSETSU speech segment to obtain acoustic features and assigned them into the terms included in it. We used the Japanese parser CaboCha [24] for detecting BUNSETSU segments.

Inter Pausal Unit(IPU) IPU is a speech segment surrounded by two pauses no shorter than 200 msec. The acoustic features were extracted from an IPU, then we equally assigned them to all the terms included in it.

Figure 2 shows an example of analysis in Japanese.

5.3. Evaluation measure

We used Mean Average Precision (MAP) [25] as an evaluation measure of retrieval performance. MAP is calculated by the mean value of the average precision of each query q in the query set Q as follows.

$$MAP = \frac{1}{|Q|}\sum_{q\in Q} AveP(q) \tag{10}$$

$AveP(q)$ represents the average precision of the list O_q of ordered retrieval results for q and is defined by Equation 11.

$$AveP(q) = \frac{1}{|R_q|}\sum_{i=1}^{O_q} \delta(O_i, R_q)\frac{\sum_{j=1}^{i}\delta(O_i, R_q)}{i} \tag{11}$$

R_q represents the relevant document set for q. $\delta(O_i, R_q)$ is a binary value indicating whether or not the output result O_i is a correct answer and is defined by Equation 12.

$$\delta(O_i, R_q) = \begin{cases} 1 & (O_i \in R_q) \\ 0 & (O_i \notin R_q) \end{cases} \tag{12}$$

5.4. Results

We constructed the prediction model using both acoustic and linguistic features. In acoustic features, “*plain set*” or “*emotion set*” were extracted in each segment(“term”, “clause”, “IPU”). That is, six prediction models were constructed. Table 3 shows the experimental results. In the columns of ”query” and ”document” in the table, the format of transcription, which is obtained either manually or by using ASR, of audio data is indicated. The column ”oracle” shows the performance when calculating $C(i, q)$ from the test data using its relevancy annotation, corresponding to the upper bound of the proposed method.

In most of the conditions, our proposed method successfully improved the retrieval performance of the baseline QLM. Among the acoustic features, the *emotion set* performed better than the *plain set*. This suggests that features used in emotion analysis are effective in analyzing spontaneously spoken query. Among the speech segments used for analysis, “term” performed best.

Next, in order to see which feature performs better than others, we repeated the experiment by using only either acoustic and linguistic features. We used “*emotion set*” and“term” for the experiments. Table 4 shows the result. It indicates that, while the linguistic features are more effective when using manual transcription, the effect decreases when using automatic transcription. Moreover, using only acoustic features performs almost equal to baseline. It also indicates that both feature sets complement each other to enhance retrieval performance.

6. Conclusions

In this work, we made better use of spontaneously spoken queries for content retrieval. We first defined *contribution for retrieval* of each term, then introduced the prediction model of it from the acoustic and linguistic features of each term in a given spoken query. The estimated contribution is used to enhance the query likelihood retrieval model to improve retrieval performance. Through our experimental evaluation, incorporation of acoustic features in a spontaneously spoken query successfully improved retrieval performance compared with the baseline QLM.

7. Acknowledgements

This work was supported by KAKENHI 16K00153.

8. References

[1] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech and Signal Processing (ICASSP), 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[2] E. Fosler-Lussier, L. Dilley, N. Tyson, and M. Pitt, "The buckeye corpus of speech: updates and enhancements." in *INTERSPEECH*, 2007, pp. 934–937.

[3] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese," in *Proc. LREC2000 (Second International Conference on Language Resources and Evaluation)*, vol. 2, 2000, pp. 947–952.

[4] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "The spoken web search task at mediaeval 2012," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8121–8125.

[5] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the ir for spoken documents task in ntcir-9 workshop," in *In Proceedings of NTCIR-9*, 2011, pp. 223–235.

[6] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, H. Nanjo, and Y. Yamashita, "Overview of the ntcir-10 spokendoc-2 task," in *In Proceedings of NTCIR-10*, 2013, pp. 573–587.

[7] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. Jones, "Overview of the ntcir-11 spokenquery&doc task," in *In Proceedings of NTCIR-11*, 2014, pp. 350–364.

[8] ——, "Overview of the ntcir-12 spokenquery&doc-2 task," in *In Proceedings of NTCIR-12*, 2016, pp. 168–179.

[9] T. Mishra, V. K. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features." in *INTERSPEECH*, 2012, pp. 1864–1867.

[10] D. N. Racca and G. J. Jones, "Incorporating prosodic prominence evidence into term weights for spoken content retrieval," in *INTERSPEECH*, 2015.

[11] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to japanese morphological analysis." in *EMNLP*, vol. 4, 2004, pp. 230–237.

[12] T. Ogiso, M. Komachi, Y. Den, and Y. Matsumoto, "Unidic for early middle japanese: a dictionary for morphological analysis of classical japanese." in *LREC*, 2012, pp. 911–915.

[13] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 275–281.

[14] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 2, pp. 179–214, 2004.

[15] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[16] B. W. Schuller, S. Steidl, A. Batliner *et al.*, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[17] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[18] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[19] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the smo algorithm for svm regression," *IEEE transactions on neural networks*, vol. 11, no. 5, pp. 1188–1193, 2000.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[21] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, 2009, pp. 131–137.

[22] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Designing an evaluation framework for spoken term detection and spoken document retrieval at the ntcir-9 spokendoc task." in *LREC*, 2012, pp. 3527–3534.

[23] T. Akiba, A. Fujii, and K. Itou, "Collecting spontaneously spoken queries for information retrieval," in *LREC*, 2004, pp. 1439–1442.

[24] T. Kudo and Y. Matsumoto, "Fast methods for kernel-based text analysis," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 24–31.

[25] E. M. Voorhees, "Variations in relevance judgments and the measurement of retrieval effectiveness," *Information processing & management*, vol. 36, no. 5, pp. 697–716, 2000.