# On the Effectiveness of Contextualisation Techniques in Spoken Query Spoken Content Retrieval

David N. Racca
ADAPT Centre
School of Computing
Dublin City University
Dublin 9, Ireland
dracca@computing.dcu.ie

Gareth J.F. Jones
ADAPT Centre
School of Computing
Dublin City University
Dublin 9, Ireland
gjones@computing.dcu.ie

## ABSTRACT

In passage and XML retrieval, contextualisation techniques seek to improve the rank of a relevant element by considering information from its surrounding elements and its container document. Recent research has demonstrated that some of these techniques are also particularly effective in spoken content retrieval tasks (SCR). However, no previous research has directly compared contextualisation techniques in an SCR setting, nor has it studied their potential to provide robustness to speech recognition errors. In this paper, we evaluate different contextualisation techniques, including a recently proposed technique based on positional language models (PLM) on the task of retrieving relevant spoken passages in response to a spoken query. We study the benefits of these techniques when queries and documents are transcribed with increasingly higher error rates. Experimental results over the Japanese NTCIR SpokenQuery&Doc collection show that combining global and local context is beneficial for SCR and that models usually benefit from using larger amounts of context in highly noisy conditions.

## 1. INTRODUCTION

Recent advances in mobile and network technologies have led to an exponential increase in the amount of spoken material that is produced and stored. Effective spoken content retrieval [9] (SCR) techniques are needed to enable information access from these type of collections. SCR has been traditionally framed as an ad-hoc retrieval task: given some information need represented by a text or spoken query, one must produce a ranked-list of spoken documents or passages in order of relevance to the query to be presented to the user for further assessment. A standard approach for SCR uses an automatic speech recognition (ASR) system to obtain a 1-best recognition transcription of the collection and standard text-based IR techniques to index the transcripts and to perform document or passage retrieval. Passage retrieval is normally preferred over full-document retrieval in cases where documents are long and multi-topical and in applications that seek to minimise audio playback time [1]. This typically requires a mechanism to identify topically coherent units of information in the speech recordings that could be treated as retrieval elements and returned in response to a query. When the spoken content is known to be delivered in a structured fashion and when structural cues are available to the SCR system, retrieval elements can be defined accordingly. Otherwise, suitable elements may be obtained by using a topic segmentation algorithm.

Although the quality of ASR systems has improved significantly over the past few years, ASR errors still pose a challenge to traditional text retrieval techniques. This occurs in domains where speech is informal, conversational, or spontaneous [9] or when the elements to be retrieved are short in length or lack sufficient contextual information and verbosity to be retrieved effectively [2]. Context and verbosity are desirable properties of an element because they can increase its chances of matching query terms, even when many of its terms are misrecognised by ASR. In general, the more repetitions of important terms used to convey the topic and the more exhaustively this topic is covered by the terms in the element to be retrieved, the more robust will be its matching process against ASR errors.

The context of an element, that is, the information that is present in its document, has been shown to be valuable for improving element-retrieval effectiveness [3]. The process of taking context into account when computing the relevance score of an element is known as contextualisation [3]. Various contextualisation techniques have been proven effective in XML retrieval [3], passage retrieval [6, 8], and SCR [12, 15] tasks. In these techniques, elements are scored depending not only on the query terms occurring within the element itself but also on those occurring in other positions within the document. Two simple and widely adopted approaches are interpolating the scores of an element with those of its document to consider global context [12], or with those from a fixed number of surrounding elements to consider local context [15]. In contrast, techniques based on positional models (PM) [6, 8] allow consideration of longer-spans of context ignoring element boundaries.

In this paper, we study the impact of incorporating context into the task of retrieving short spoken passages given a spoken query from a collection of long spoken documents. Our hypothesis is that context becomes increasingly valuable for improving retrieval effectiveness as ASR errors increase in the transcripts of the spoken queries and documents. We

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*SIGIR '16, July 17-21, 2016, Pisa, Italy*
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07. . . $15.00
DOI: http://dx.doi.org/10.1145/2911451.2914730

**Table 1: List of manual and ASR transcript types.**

| SpokenQuery&Doc ID | Short ID |
|---|---|
| MANUAL | M |
| K-REF-WORD-MATCH | A0 |
| REF-WORD-MATCH | A1 |
| REF-WORD-UNMATCHLM | A2 |
| REF-WORD-UNMATCHAMLM | A3 |

validate this empirically by comparing the performance of various contextualisation techniques in a SCR task.

The remainder of the paper is organised as follows. Section 2 describes the spoken test collection. Section 3 presents our baseline retrieval approach which we extend with the contextualisation techniques presented in Section 4. Section 5 describes experiments and results obtained. Finally, Section 6 concludes and discusses future work.

## 2. SPOKEN TEST COLLECTION

The SDPWS dataset has been used at recent editions of the NTCIR SpokenQuery&Doc (SQD) tasks [1]. This corpus contains speech recordings of 98 oral presentations in Japanese totalling approximately 27 hours. The task organisers provided automatic transcripts of the recordings of varying quality. These were generated using an ASR system with different combinations of acoustic and language models. Furthermore, human transcripts are provided which are manually annotated with the times when slide transitions were made by the presenters, as well as groups of consecutive slides used to present a single topic or idea. These slide groups define a list of topical segments within a presentation, used in the task as pre-defined passages to be retrieved in response to a query. In our experiments, we similarly use these segments as retrieval passages. Table 1 lists the different transcript types. The first column shows the ID of each type as mentioned in the task overview [1], while the second shows short IDs used throughout this paper.

Since no characters are placed between words in written Japanese, we process text transcripts with the morphological analyser MeCab[1] to obtain tokens that can be used as indexing features. For this purpose, we configure MeCab to output the base form of the identified words along with their part-of-speech tags. As lemmas of nouns and verbs are effective indexing features for Japanese SCR [12], we remove all tokens not tagged as verbs or nouns. Further, we remove function words that are misclassified as nouns or verbs with a stop word list containing 44 of the most frequent prepositions and determiners. After processing the text, we use the Terrier platform v4.0[2] to generate an index with positional information for each transcription type. Table 2 shows term statistics of post-processed documents and passages. In total, the collection comprises 98 documents which are split into 2330 passages. For each transcript, we report word error rates (WER) averaged across utterances and term error rates (TER) [7] averaged across passages.

**Table 2: Collection statistics of documents and passages.**

| ID | WER | TER | #Terms | Documents | | Passages | |
|---|---|---|---|---|---|---|---|
| | | | | Ave.len. | S.D.len. | Ave.len. | S.D.len. |
| M | 0% | 0% | 6230 | 1769.17 | 276.15 | 74.48 | 67.61 |
| A0 | 22.0% | 39.3% | 6350 | 1763.09 | 274.01 | 74.19 | 67.42 |
| A1 | 43.7% | 70.0% | 6131 | 1752.15 | 262.81 | 73.62 | 67.56 |
| A2 | 67.5% | 121.9% | 11219 | 1922.76 | 285.21 | 80.73 | 74.77 |
| A3 | 70.5% | 120.6% | 14190 | 1594.50 | 247.70 | 67.06 | 62.55 |

**Table 3: Error rates and statistics of query transcripts.**

| Queries | ID | WER | TER | #Terms | Ave.len. | S.D.len. |
|---|---|---|---|---|---|---|
| SQD-1 | M | 0% | 0% | 373 | 24.13 | 11.61 |
| | A1 | 51.6% | 80.6% | 490 | 29.64 | 15.14 |
| | A2 | 77.7% | 125.5% | 871 | 39.10 | 23.34 |
| | A3 | 69.1% | 109.6% | 754 | 32.89 | 19.64 |
| SQD-2 | M | 0% | 0% | 714 | 30.77 | 12.06 |
| | A0 | 33.7% | 45.3% | 766 | 30.32 | 13.67 |
| | A1 | 49.2% | 68.3% | 961 | 36.85 | 16.65 |
| | A2 | 75.4% | 111.8% | 1763 | 50.81 | 28.37 |
| | A3 | 66.1% | 98.9% | 1615 | 42.41 | 21.09 |

The spoken queries used in the SQD-1 and SQD-2 tasks are available for the SDPWS collection. These sets contain 35 and 80 queries respectively. Also, organisers generated human and ASR transcripts for the queries by using the same ASR models used to transcribe the presentations. Table 3 summarises term statistics and ASR error rates for the spoken queries. Relevance judgements for these query sets were created by the task organisers from a pool of results submitted for the SQD tasks. Three relevance levels were annotated: full, partial, and no relevance. In our experiments, we treat partially relevant passages as fully relevant.

## 3. BASIC RETRIEVAL MODEL

In our simplest retrieval set-up, we treat passages as retrieval elements and use standard document retrieval to rank them in order of their relevance to the query. Our ranking function is based on the Okapi BM25 function of probabilistic retrieval [14]. Besides the well known $k_1$, $b$, and $k_3$ parameters, we include a fourth parameter, namely $d \geq 1$, as the exponent of the IDF weight [16]. This parameter can be adjusted to increase the relative difference between weights assigned to frequent and rare terms. In our experiments, setting $d > 1$ results in improved effectiveness as this may provide better estimates of IDF weights for small collections. Similar effects can be obtained with the modification proposed in [11], although this results in lower retrieval effectiveness in our task. The resulting function calculates the weight of a term $t$ occurring in element $e$ and query $Q$ as:

$$w_e(t) = \frac{(k_1+1)tf(t)}{tf(t)+k_1(1-b+b\frac{\mathrm{docl}}{\mathrm{avel}})} \frac{(k_3+1)qf(t)}{qf(t)+k_3} w_1(t)^d \quad (1)$$

where $tf(t)$ and $qf(t)$ denote the number of occurrences of $t$ in $e$ and $Q$ respectively, *docl* is the length of $e$, *avel* is the average length of all the elements of the same type in the collection, $n_t$ is the number of elements containing term $t$, $N$ is the number of elements in the collection, $w_1(t)$ is $\log(N - n_t + 0.5) - \log(n_t + 0.5)$ and $b$, $k_1$, $k_3$, and $d$ are tuning parameters. Given this term scoring function, we rank elements according to their relevance score with respect to $Q$ given by $S_{BM25}(Q, e) = \sum_{t \in Q \cap e} w_e(t)$.

## 4. CONTEXTUALISATION TECHNIQUES

In this section, we present the contextualisation techniques that are investigated in our study.

[1] http://mecab.sourceforge.net/

[2] http://terrier.org

## 4.1 Document Score Interpolation

A simple and popular contextualisation approach consists of combining a passage relevance score with the score of its document. Firstly, passages and documents are scored independently to form two separate ranked-lists of results. Secondly, initially retrieved passages are re-ranked according to the combination of their document scores. For score combination, we adopt a simple weighted linear sum of scores or CombSUM [5]. The relevance score of passage $p$ within document $D$ is given by:

$$S_{DSI}(Q,p) = \lambda S_{BM25}(Q,D) + (1-\lambda) S_{BM25}(Q,p) \quad (2)$$

where the interpolation parameter $\lambda$ adjusts the influence of the document score over the combined score.

## 4.2 Positional Models

Positional models (PM) seek to improve IR effectiveness by exploiting information about the positions where query terms occur in a document. A representative example of these models are positional language models (PLM) [10] which were introduced in IR as a mechanism to integrate evidence from term proximity features and passages into the language modelling framework. A PLM estimates the probability $P(t|i,D)$ that term $t$ is generated at position $i$ in document $D$. In this estimation, the so-called pseudo-frequency of the term is used which is calculated by means of a kernel decay function that propagates occurrences of the term to distant positions in the document. This probability can possibly be influenced by all occurrences of $t$ in $D$ depending on their distance to $i$. The kernel function is commonly parametrised by a propagation parameter $\sigma$ which adjusts the influence that a term occurrence has over distant positions. Among these, the Gaussian kernel $\exp(-(j-i)^2/2\sigma^2)$ has been shown effective in previous studies [10, 6, 8].

In recent work, PMs were proposed as a contextualisation technique for passage retrieval [6]. In this work, a standard TFIDF approach was used to compute the relevance score of a passage $p$ within document $D$ where the frequency of term $t$ in $p$ is given by its pseudo-frequency estimate:

$$tf_{PM}(t) = \sum_{i \in pos(t,D)} \sum_{j=p_1}^{p_n} kernel(j,i) \quad (3)$$

where $pos(t,D)$ is the set of positions where $t$ occurs in $D$, $p_1, \ldots, p_n$ are the spanning positions of $p$ in $D$, and $kernel$ is the symmetric Gaussian kernel. To avoid longer passages from receiving unmerited pseudo-frequency counts, the summation over the kernel function is applied across a fixed number of positions independent of $p$'s length.

In our experiments with PMs, we incorporate the pseudo-frequencies into the TF factor within the Okapi BM25 model. This is performed by replacing the raw term frequency $tf(t)$ by $tf_{PM}(t)$ in Equation 1 to obtain a modified weight $w'_e$. The resulting passage scoring function is then defined as:

$$S_{PM}(Q,p) = \sum_{t \in Q \cap D} w'_p(t) \quad (4)$$

To reduce the execution time of our scoring algorithm, we evaluate the kernel only at the position within the passage that gives its maximum value. That is, we evaluate the kernel at $j = p_1$ or $j = p_n$ if $i < p_1$ or $i > p_n$ respectively, or at $j = i$ otherwise. Finally, we also experiment with a model that combines passage scores obtained with $S_{PM}$ and document scores obtained with $S_{BM25}$, this is:

$$S_{DSI\text{-}PM}(Q,p) = \lambda S_{BM25}(Q,D) + (1-\lambda) S_{PM}(Q,p) \quad (5)$$

# 5. RETRIEVAL EXPERIMENTS

To study the value of context for improving passage ranking in noisy conditions, we evaluate the effectiveness of the models presented in Sections 3 and 4 on various combinations of query and document transcripts. In all experiments, retrieval effectiveness is measured in terms of mean average precision (MAP) at depth 1000. Each combination of transcripts imposes a different evaluation condition with varying level of noise, each of which may require adjusting model parameters differently in order to achieve optimal performance. Further, in order to validate our hypothesis, we would like to find values for the contextualisation parameters $\sigma$ and $\lambda$ that provide the best performance in each case. Consequently, we optimise parameters for each model by seeking to maximise retrieval effectiveness in each evaluation condition. The optimisation method we use is a greedy iterative approach called "Promising Directions" [13] which performs multiple line searches over decreasing trusted regions in the parameter search space. Since MAP is not a smooth nor a convex function [13] this method is not guaranteed to find a global maximum. Nevertheless, it can still find configurations that perform significantly better in our task than if using default BM25 parameters[3].

To obtain a fair estimate of the relative performance of the models, we optimise parameters on the SQD-1 queries and evaluate on the SQD-2 queries. Table 4 reports MAP scores obtained by the BM25, DSI, PM, and DSI-PM models on the SQD-2 queries for different combinations of transcript types. The percentages next to the MAP scores show relative differences with respect to BM25. In each row, bold values and markers *, †, and ⋄ indicate statistically significant differences with respect to BM25, DIS, PM, and DIS-PM respectively based on a $MaxT$ permutation test ($B = 100,000$, $\alpha = 0.05$) that corrects for multiple hypothesis testing [4]. Also, we report query-averaged TERs for each transcript combination calculated by restricting terms to only those occurring in the queries. Overall, results indicate that using global (DSI) and local (PM) context either in isolation or in combination (DSI-PM) provide gains in retrieval effectiveness across all evaluation conditions. Furthermore, the relative gains of using context in highly noisy conditions (TER $> 55\%$) are greater on average than in less noisy conditions.

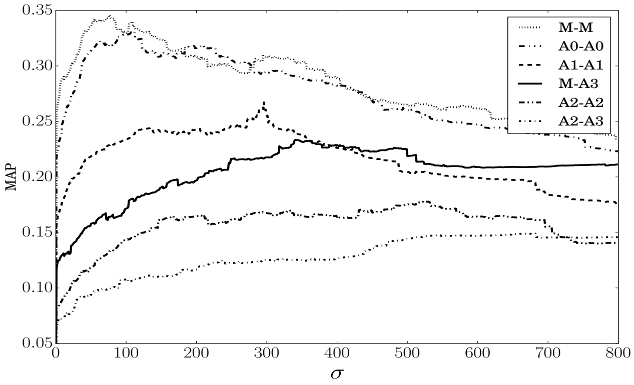
Finally, we study the effects of varying the contextualisation parameter $\sigma$ in the PM scoring function. Figure 1 shows how MAP scores vary as we increase the values of $\sigma$ in six representative evaluation conditions. Data points were generated considering optimal parameter values for the SQD-2 queries. For perfect or quasi-perfect transcripts (M-M and A0-A0), the model achieves maximum performance at $\sigma = 76$ and $\sigma = 111$ respectively, while for moderately noisy transcripts (A1-A1 and M-A3) it does so at $\sigma = 296$ and $\sigma = 341$ respectively. Finally, in extremely noisy conditions (A2-A2 and A2-A3), the maximum points are located at $\sigma = 530$ and $\sigma = 682$ respectively. These observations provide supporting evidence for our claim that longer spans

[3] Due to space limitations, we do not report here the full set of optimal parameter values which generally vary across models and transcripts. These are available at: github.com/dracca/context.

**Table 4: MAP scores for test queries.**

| TER | Transcripts | | Models | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Query | Doc. | BM25 | PM | | DSI | | DSI-PM | |
| 0% | M | M | .272 | .291 | +6% | **.305** | +12% | .314 | +15% |
| 20% | M | A0 | .261 | .294 | +12% | .260 | 0% | .299 | +14% |
| 43% | M | A1 | .228 | .267 | +16% | **.272** | +19% | .278 | +21% |
| 51% | A0 | A0 | .261 | .292 | +11% | .261 | 0% | .293 | +12% |
| 58% | M | A2 | .085 | **.178** | +108% | **.160** | +87% | **.174** | +103% |
| 62% | A0 | A1 | .236 | .267 | +13% | **.271** | +14% | .274 | +15% |
| 71% | A0 | A2 | .091 | **.177** | +94% | **.166** | +82% | **.172** | +88% |
| 78% | M | A3 | .119 | **.209** | +75% | **.214** | +80% | **.192** | +60% |
| 82% | A1 | A1 | .186 | **.239** | +28% | **.249** | +34% | **.277** | +49% |
| 87% | A1 | A2 | .097 | .124 | +27% | **.126** | +29% | .139 | +43% |
| 88% | A0 | A3 | .111 | **.173** | +55% | **.191** | +72% | **.167** | +50% |
| 100% | A2 | A2 | .117 | .096 | -21% | .112 | -4% | .155† | +33% |
| 106% | A1 | A3 | .048 | **.144**** | +200% | **.142**◇ | +196% | **.134** | +178% |
| 108% | A2 | A3 | .066 | .100 | +52% | **.122** | +86% | .126 | +91% |
| 115% | A3 | A3 | .095 | .145 | +52% | **.146** | +53% | **.168** | +76% |

**Figure 1: MAP scores versus $\sigma$ for the PM model.**



of context become increasingly useful for retrieval as ASR errors increase in the transcripts.

# 6. CONCLUSIONS

This paper has investigated the benefits of using context for improving the ranking of spoken passages given a spoken query. Three contextualisation techniques have been evaluated and compared against a well-tuned state-of-the-art retrieval model: a document score interpolation, a positional model, and their combination. Results of retrieval experiments with transcripts of varying quality validate previous findings that highlight the importance of using context in element-retrieval and SCR tasks and indicate that a combination of local and global context performs best for SCR. Further analysis reveals that considering greater extents of local context can improve IR effectiveness as ASR errors increase in the transcripts. In future work, we will evaluate models on an unsegmented spoken collection where we could fully exploit the "soft" characterisation of passage boundaries provided by PMs. Since recognition quality may vary greatly across utterances, we also plan to develop new techniques that could re-adjust context-incidence parameters according to ASR confidence estimates.

# 7. ACKNOWLEDGEMENT

This work is supported by Science Foundation Ireland through the CNGL Programme (Grant No: 12/CE/I2267) in the ADAPT Centre at Dublin City University.

# 8. REFERENCES

[1] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones. Overview of the NTCIR-12 SpokenQuery&Doc-2 task. In *Proceedings of NTCIR-12*, Tokyo, Japan, 2016.

[2] J. Allan. Perspectives on information retrieval and speech. In *Information Retrieval Techniques for Speech Applications*, pages 323–326. 2002.

[3] P. Arvola, J. Kekäläinen, and M. Junkkari. Contextualization models for XML retrieval. *Information Processing & Management*, 47(5):762–776, 2011.

[4] L. Boytsov, A. Belova, and P. Westfall. Deciding on an adjustment for multiplicity in IR experiments. In *Proceedings of SIGIR'13*, pages 403–412, Dublin, Ireland, 2013.

[5] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of SIGIR'94*, pages 302–310, Dublin, Ireland, 1994.

[6] D. Carmel, A. Shtok, and O. Kurland. Position-based contextualization for passage retrieval. In *Proceedings of CIKM'13*, pages 1241–1244, San Francisco, CA, USA, 2013.

[7] S. E. Johnson, P. Jourlin, K. S. Jones, and P. C. Woodland. Spoken document retrieval for TREC-7 at Cambridge University. In *Proceedings of TREC-7*, pages 191–200, 1999.

[8] M. Keikha, J. H. Park, W. B. Croft, and M. Sanderson. Retrieving passages and finding answers. In *Proceedings of ADCS'14*, pages 81–84, Melbourne, Australia, 2014.

[9] M. Larson and G. J. F. Jones. Spoken content retrieval: A survey of techniques and technologies. *Foundations and Trends in Information Retrieval*, 5(4—5):235–422, 2012.

[10] Y. Lv and C. Zhai. Positional language models for information retrieval. In *Proceedings of SIGIR'09*, pages 299–306, Boston, MA, USA, 2009.

[11] M. Murata, H. Nagano, R. Mukai, K. Kashino, and S. Satoh. BM25 with exponential IDF for instance search. *IEEE Transactions on Multimedia*, 16(6):1690–1699, 2014.

[12] H. Nanjo, T. Yoshimi, S. Maeda, and T. Nishio. Spoken document retrieval experiments for SpokenQuery&Doc at Ryukoku University (RYSDT). In *Proceedings of NTCIR-11*, pages 365–370, Tokyo, Japan, 2014.

[13] S. Robertson and H. Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[14] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *Proceedings of TREC-3*, pages 109–126, 1995.

[15] S.-R. Shiang, P.-W. Chou, and L.-C. Yu. Spoken term detection and spoken content retrieval: Evaluations on NTCIR 11 SpokenQuery&Doc task. In *Proceedings of NTCIR-11*, pages 371–375, Tokyo, Japan, 2014.

[16] C. Zhai. Notes on the Lemur TFIDF model. Technical report, Carnegie Mellon University, 2001.