Incorporating Prosodic Prominence Evidence into Term Weights for Spoken Content Retrieval

David N. Racca, Gareth J.F. Jones

ADAPT Centre
School of Computing
Dublin City University
Dublin 9, Ireland

dracca@computing.dcu.ie, gjones@computing.dcu.ie

10.21437/Interspeech.2015-52

Abstract

We present an extended technique for spoken content retrieval (SCR) that exploits the prosodic characteristics of spoken terms in order to improve retrieval effectiveness. Our method promotes the rank of speech segments containing a high number of prosodically prominent terms. Given a set of queries and examples of relevant speech segments, we train a classifier to learn differences in the prosodic realisation of spoken terms mentioned in relevant and non-relevant segments. The classifier is trained with a set of lexical and prosodic features that capture local variations of prosodic prominence. For an unseen query, we perform SCR by using an extension of the Okapi BM25 function of probabilistic retrieval that incorporates the prosodic classifier's predictions into the computation of term weights. Experiments with the speech data from the SDPWS corpus of Japanese oral presentations, and the queries and relevance assessment data from the NTCIR SpokenDoc task show that our approach provides improvements over purely text-based SCR approaches.

1. Introduction

Fully realising the potential of large collections of digital spoken content requires an effective integration of automatic speech recognition (ASR) and information retrieval (IR) technologies [1]. The field of spoken content retrieval (SCR) investigates how to best combine these two technologies in order to enable users to find relevant information.

Prosody, that is, the rhythm, intonation, and stress patterns of speech, plays an important role in human-to-human communication [2, 3]. For instance, prosodic phrasing, or grouping, is sometimes used to disambiguate among syntactic structures, and to facilitate the understarding of utterances in general [2, 3]. Moreover, speakers may pronounce some words more saliently than others. This phenomenon is commonly referred to as prosodic prominence, and is believed to encode the informational status of words and how this status changes over time. In this regard, there is evidence that spoken words considered "new", "important", "focussed", "not given", "unpredictable", or "inaccessible" in a discourse are more likely to be made prominent than others [2, 3].

The method proposed in this paper is based on the hypothesis that prosodic prominence can be used as a cue to estimate the importance of words in specific portions of spoken content. In the context of IR, a word is considered important or informative for a document if the word is significantly associated with the topic of the document and if it is effective in discriminating this topic from others. If it is true that the most prominent words are those that best describe the topic in discourse, an SCR system could then exploit this to generate better estimates of the importance, or weight, that a word is given in a particular portion of speech. In other words, words that are made prominent could be considered more representative of the topic of the segment, and hence given increased emphasis in the SCR process.

In this work, we introduce a retrieval method designed to exploit the relative prosodic prominence of terms spoken in relevant and non-relevant portions of speech. First, given a set of queries for which we have relevance judgements, i.e., we know which speech segments are relevant to each query, we train a classifier to distinguish between query terms appearing in relevant segments from those not appearing in relevant segments. The classifier is only trained with speaker-normalised prosodic and lexical features that are known to be useful for the tasks of pitch accent and prominence detection. Our hypothesis is that this classifier will learn the relationship between prominence and importance of significant terms. Predictions from this prosodic model are then incorporated into a retrieval function that ranks speech content according to their probability of relevance to a query. This function takes into account frequency statistics about the distribution of terms in the collection plus statistics from the labels predicted by the classifier.

This paper is organised as follows. Section 2 presents previous and related work. Section 3 describes our prosodic weighting technique in more detail. Section 4 presents experimental results. Finally, Section 5 reports our conclusions and points out directions for future research.

2. Related work

This section reviews existing term weighting techniques and overviews relevant previous research that has investigated the usefulness of prosodic information in SCR tasks.

2.1. Term weighting techniques for SCR

The task of an SCR system is to retrieve content from a collection of speech documents that is relevant to the searcher's information need, in response to the user's query q. To fulfill this goal, an ASR is first used to obtain text transcripts for the spoken content. Transcripts are then split into smaller content passages or segments using a topic segmentation algorithm in order to obtain topically homogeneous segments. Additionally, frequent words are usually removed from the segments and only the stems (root) form of words are considered. In this paper, we

refer to word stems as indexing terms.

Segments are indexed by using text-based IR techniques. Two widely used IR models are the probabilistic approach [4] and the vector space model [5]. These IR models rank retrieved segments according to an estimate of their relevance with respect to q, known as the retrieval status value (RSV). Commonly, the RSV for a segment s and query q is defined as shown in Equation 1, where the sum ranges across all the terms t that occur in both q and s, and $w(t)$ denotes the individual contribution weight of each term t.

$$RSV(q,s) = \sum_{t \in q,s} w(t) \tag{1}$$

In the probabilistic approach, a popular model is the Okapi BM25 function [6] in which the term weights $w(t)$ are calculated as the product of a term frequency (TF) function and an inverse document frequency (IDF) factor, as depicted in Equation 2.

$$w(t) = \frac{(k_1+1)\,\mathrm{tf}_t}{\mathrm{tf}_t + k_1(1-b+b\frac{\mathrm{docl}}{\mathrm{avel}})} . \log \frac{N - n_t + 0.5}{n_t + 0.5} \tag{2}$$

In the TF factor, tf_t denotes the number of times that t occurs in s, docl denotes the segment length in number of terms of s, avel denotes the segments' average length in the collection, and k_1 and b are scalar parameters. In the IDF factor, N and n_t denote the total number of segments in the collection and the number of segments containing t respectively.

2.2. Prominent vs. informative words

In previous work, Silipo and Crestani investigated the relationship between prosodic prominent and informative words in the English subset of the OGI Stories Corpus [7, 8]. Specifically, they compared manually-annotated acoustic stress levels against weights computed with the Okapi BM25 function for every term in the corpus. The authors reported that there exists some correlation, though not significant, between the stress levels of terms and their BM25 weights.

2.3. Simple prosodic term weighting techniques

In [9, 10, 11, 12], the authors experiment with various methods that exploit the prosodic prominence of terms to improve retrieval effectiveness in different speech retrieval tasks.

First, every term mention within a speech segment is assigned an acoustic score. This score is generated from a combination of prosodic features extracted from the speech signal, and is assumed to reflect the grade of relative salience of the term mention in the context where it is pronounced. Among the most common prosodic features used for this purpose are intensity [9, 10, 11, 12], pitch [10, 11, 12], duration [9, 11, 12], and ASR confidence scores [9]. Next, the acoustic scores from term mentions are combined into a single acoustic score for each unique term in a segment. To achieve this, the scores are usually averaged [9, 10] or only their maximum value is retained [10, 11, 12].

Second, the acoustic scores are incorporated into the computation of the weights $w(t)$ within the RSV function. Different term weighting functions have been proposed. Chen et al. [9] and Guinaudeau and Hirschberg [10], use a vector space model (VSM) [5] to represent segments and the cosine similarity as the RSV function. Chen et al. compute term weights with the function shown in Equation 3 where $\mathrm{ac}(t,k)$ denotes the acoustic score of the k-th mention of the term t in segment s. Guinaudeau and Hirschberg use the function depicted in Equation 4 where $\mathrm{ac}(t)$ denotes the acoustic score assigned to t in s, θ_{ir}, θ_{ac} are tuning parameters, and $\text{tf-idf}_l(t)$ is Lecorvé et al.'s TF-IDF function [13].

$$w(t) = (1 + \log(\sum_k \mathrm{ac}(t,k))) . \log(N/n_t) \tag{3}$$

$$w(t) = \frac{\theta_{\mathrm{ir}} . \text{tf-idf}_l(t) + \theta_{\mathrm{ac}} . \mathrm{ac}(t)}{\theta_{\mathrm{ir}} + \theta_{\mathrm{ac}}} \tag{4}$$

In previous work [11, 12], we experimented with the functions shown in Equations 4 and 5. In the former, we replaced Lecorvé et al.'s TF-IDF by the TF-IDF function depicted in Equation 6. The later is a weighted linear interpolation between $\mathrm{ac}(t)$ and the term's TF factor calculated with Equation 6.

$$w(t) = \mathrm{idf(t)} .(\alpha . \mathrm{tf}(t) + (1-\alpha)\,\mathrm{ac}(t)) \tag{5}$$

$$\text{tf-idf}(t) = \frac{k_1 . \mathrm{tf}_t}{\mathrm{tf}_t + k_1\left(1 - b + b\frac{\mathrm{docl}}{\mathrm{avel}}\right)} . \log(N/n_t + 1) \tag{6}$$

A range of issues can be identified with this existing work. First, acoustic scores are computed in an ad-hoc manner by calculating aggregated statistics over the pitch and intensity contours, which are then combined linearly across term occurrences. This is not ideal, since valuable information about the shape of the contours may be lost in the process, as well as any interdependency that may exist between features. Second, acoustic scores are extempore integrated into term weighting functions, possibly breaking theoretically well-founded retrieval models, which could lead to suboptimal retrieval performance [14]. Finally, no significant improvements have been reported in SCR tasks so far [9, 11, 12].

The technique proposed in this paper addresses the first issue by training a classifier with an extended feature-set that, besides aggregations, includes other features developed recently by [15, 16]. Moreover, the classifier is not trained with annotations of prominence, nor is it used to predict intermediate prominence levels, but is instead trained to predict a term's significance by using annotations of relevance. Despite this approach requiring manual relevance judgmenents that are expensive to obtain, it provides a more direct model of the relationship between prosody and relevance. The second issue is addressed with a weighting function that, in constrast to Equations 4 and 5, preserves the non-linear saturation of the TF factor and, in contrast to Equation 3, is more sensitive to the acoustic scores.

3. Prosodic term weighting for SCR

This section presents our prosodic term weighting approach.

3.1. Prosodic-relevance dataset creation

3.1.1. Features

For each term occurrence, we extract 294 prosodic features. This feature set includes Rosenberg's features [15] for pitch accent detection, which are an extension of the features previously proposed by Mishra et al. in [16] for word prominence detection. The AuToBI toolkit v1.5.1 [17] is used for feature-extraction since it implements all the necessary algorithms. We also include raw duration extracted from the forced alignment output of an ASR system, plus a normalised version of duration based on the number of pseudo-syllables identified across the target term's region. The length of silences preceding and

following each term is also included since it is considered useful for the task [18]. To capture local variations of prominence around a term occurrence, features are extracted from local context windows around the term's region. We used the 8 context windows used in [15], which include zero, one, or up to two of the terms preceding and following the target term. In addition, we include the part-of-speech (POS) tag of the term as well as the POS tags of the two terms preceding and following the target term.

3.1.2. Target class

Let C be the collection of segments $s_1, \ldots, s_J$, t a particular term occurring in C and t_{kj} its k-th occurrence in s_j. Let Q be a set of queries $q_1, \ldots, q_L$ and $R_1, \ldots, R_L$ their respective relevance assessment sets such that R_l contains the segments that are relevant to q_l. Additionally, we write $t \in q_l$ to say that t occurs in the query q_l. Every term occurrence t_{kj} can be labelled depending on whether t appears in a query q_l for which the segment s_j is relevant. Formally, we label every term occurrence t_{kj} as relevant according to Equation 7.

$$R(t_{kj}) \Leftrightarrow \exists l : t \in q_l \wedge s_j \in R_l \tag{7}$$

3.2. Training of prosodic-relevance models

We create a training set by selectively removing some non-relevant term instances. The main motivation behind this is that, when constructing relevance judgements for evaluating IR systems, usually only a small proportion of the collection set can be manually assessed for relevance. Therefore, it is likely that some segments that are indeed relevant (or non-relevant) to a query were actually never checked by a human-assessor and remain unlabelled. First, we remove occurrences of terms that do not appear in any of the queries in Q since these do not provide any examples of relevant instances. Second, if some segments are known to be non-relevant or known not to be judged, then the instances corresponding to these segments can be removed from the training set. We further remove all the occurrences of terms that do not fall in the noun or verb part-of-speech category, plus those occurrences of terms that appear in more than half of the segments in the collection. Highly frequent terms are mainly discarded to avoid generating bias in the training set towards less informative terms.

The LIBSVM toolkit v3.18 [19] is used to train support vector machines (SVMs) using a radial basis kernel function (RBF). We have previously experimented with classification trees, linear SVMs, and multilayer perceptrons but these did not outperform RBF-based SVMs. In order to find an SVM model that works well on unseen data, we use grid search and cross validation to find the best values for the C and γ parameters. Moreover, as the number of relevant instances is normally much less than the number of non-relevant instances, we use balanced accuracy (BAC) [20] to evaluate the generalisation power of the models when performing cross validation. We also set different cost weights for the minority (relevant) and majority (non-relevant) classes. We set the cost weight of the majority class to 1 and the cost weight of the minority class to the ratio between the majority and minority classes. Once a good SVM model is found, it is used to predict labels for all the term occurrences in the collection, including those that have been discarded in the construction of the training set. In the rest of the paper, we use $L_D(t_{kj})$ to denote the labelling function of the prosodic-relevance model obtained from dataset D. The function $L_D(t_{kj})$ outputs 1 whenever t_{kj} is relevant according to the model trained with D and 0 otherwise.

3.3. Indexing and retrieval

The Terrier platform v3.5[1] [21] is used to perform indexing and retrieval. Terrier provides out-of-the-box implementations of IR models and weighting functions, including the Okapi BM25. When building the index, Terrier stores statistics about the distributions of terms in the collection. We extend Terrier to also store frequency statistics about the predictions made by the prosodic-relevance model.

Retrieval is performed by using the Okapi BM25 weighting function. In order to incorporate the prosodic-model's predictions into the BM25 function, we follow the method previously proposed by Song et al. in [22] for incorporating term proximity measures. Instead of incorporating evidence from individual term occurrences as an external factor into the TF-IDF multiplication, Song et al. replace the raw term frequency tf_t by a function $\mathrm{TF}(t)$ that combines evidence from the individual occurrences of the term t in the segment. This maintains the property of non-linearity of the term frequency factor.

Equation 8 presents our proposed definition for $\mathrm{TF}(t)$. This incorporates the predictions of a prosodic-relevance model trained with dataset D, where $f_D(t,k)$ is defined as shown in Equation 9.

$$\mathrm{TF}(t) = \sum_k f_D(t,k) \tag{8}$$

$$f_D(t,k) = \begin{cases} 1 & \text{if } n_t \geq N/2 \\ \alpha & \text{if } n_t < N/2 \ \wedge \ L_D(t_{kj}) = 1 \\ \alpha^{-1} & \text{otherwise} \end{cases} \tag{9}$$

4. Retrieval experiments

The retrieval effectiveness of our prosodic weighting technique is evaluated by running SCR experiments over a collection of audio recordings. In this section, we describe the test collection used and present experimental results.

4.1. Speech collection

We used the speech recordings from the 1st to the 7th Spoken Document Processing Workshop (SDPWS) corpus. This dataset contains 98 recordings of academic presentations in Japanese, comprising 27 hours of speech data. Manual and ASR transcripts are available for each presentation in the dataset, as well as annotations about the times when slide transitions were made by the presenters. Moreover, groups of slides used to describe a single topic are available, and we used these to segment the transcripts into topically cohesive units. Japanese text is mainly processed as described in [12], but with two major differences. First, the text was tokenised with the Japanese morphological analyser MeCab[2] instead of ChaSen[3]. Second, we removed terms that are not tagged as nouns or verbs by MeCab. By pre-processing transcripts in this way, we obtained the best retrieval results for our baseline runs.

4.2. Query sets and prosodic-relevance datasets

We used the queries and relevance assessment data from the NTCIR-10 SpokenDoc-2 [23] and the NTCIR-11 SpokenQuery&Doc [24] tasks. These two query sets were formulated

[1] `http://terrier.org`
[2] `http://mecab.sourceforge.net/`
[3] `http://chasen-legacy.sourceforge.jp`

for the SDPWS collection. While the SpokenDoc-2 query set (SD2) includes 120 text queries [23], the SpokenQuery&Doc query set (SQD) contains 37 spontaneous spoken queries [24]. Manual and ASR transcripts are available for spoken queries. In our experiments, we only used the manual transcripts of the spoken queries and processed them in the same way as the presentation transcripts.

Following the procedure described in Section 3.1, we generated four prosodic-relevance datasets, one for each combination of transcripts (MAN, ASR) and query sets (SD2, SQD). Next, by using the method described in Section 3.2, we trained a prosodic-relevance model with each of the prosodic-relevance datasets, from which we obtained the labelling functions $L_{\text{MAN-SD2}}$, $L_{\text{MAN-SQD}}$, $L_{\text{ASR-SD2}}$, and $L_{\text{ASR-SQD}}$. These models obtained a cross-validation BAC of 0.64, 0.60, 0.63, and 0.59 respectively. These values drop to about 0.54 when the models are evaluated across datasets.

4.3. Retrieval experiments

We compare the retrieval performance of three retrieval models: the text-based Okapi BM25 model presented in Section 2.1 (B), Guinaudeau's harmonic mean with Okapi BM25 weighting (G), and our extension of BM25 presented in Section 3.3 (L) which integrates predictions from one of the prosodic-relevance models $f_D(t,k)$ introduced in Section 4.2. Guinaudeau's approach was implemented as described in [12] with acoustic scores defined as the product between the maximum values of RMS energy and F0 across term occurrences.

We ran four retrieval experiments in total, one for each combination of text transcripts and query sets. In experiment MAN-SD2-SQD, we indexed the manual transcripts of segments, used the function $f_{\text{MAN-SD2}}$ for retrieval, optimised parameters with the SD2 query set, and evaluated over the SQD query set. In experiment MAN-SQD-SD2, retrieval was performed with $f_{\text{MAN-SQD}}$, the parameters were tuned with the SQD query set, and the evaluation was performed over the SD2 query set. Similarly, ASR-SD2-SQD and ASR-SQD-SD2 identified the same class of experiments when indexing the ASR collection of segments and performing retrieval with the $f_{\text{ASR-SD2}}$, $f_{\text{ASR-SQD}}$ functions, respectively. For the baseline and prosodic retrieval models, we explored different values for the parameters k_1, b, θ_{ir}, θ_{ac}, and α. More precisely, we experimented with $0 \leq k_1 \leq 6$, $0 \leq b \leq 1$, $1 \leq \theta_{ir}, \theta_{ac} \leq 5$ and $\alpha = 2, 4, 8, 16, 32$. Retrieval effectiveness was evaluated in terms of mean average precision (MAP).

4.4. Results

Table 1 shows two groups of results for each experiment. The first group depicts the results obtained by the retrieval models when the parameters are optimised with the query set shown in the "Train" column. The second group shows the best results obtained by the models when the optimal configuration of parameters is used. In each group, MAP values in bold indicate statistically significant improvements over the worse result based on paired t-tests with 95% confidence.

The results indicate that our prosodic model significantly outperforms the B and G models in the MAN-SD2-SQD experiment. On the contrary, our model significantly underperforms the models B and G in the MAN-SQD-SD2 experiment. The exact reasons for this are unclear and need further analysis. We hypothesise that two factors affect the performance of our model in the MAN-SQD-SD2 experiment. First, the cross-validation BAC obtained by the $L_{MAN-SQD}$ model is 6% worse than the BAC obtained by the $L_{MAN-SD2}$ model. Second, as can be seen from the second group of results, the optimal configuration of parameters differs substantially from one query set to the other. The higher values of k_1 in the optimal models for SQD reflect that there is greater room for improvement on this query set by exploiting within-document term frequencies. The improvements observed in MAN-SD2-SQD are not present in the experiments with ASR transcripts, however the general trend is that L performs better than the rest of the models. Furthermore, the results reveal that Guinaudeau's model performs the same as the text-based Okapi BM25 model in this task.

Table 1: *Retrieval results for the prosodic and baseline models.*

Trans.	Query Set		Model	Parameters					MAP
	Train	Test		k_1	b	α	θ_{ir}	θ_{ac}	
MAN	SD2	SQD	L	0.8	0.2	8	-	-	**.200**
			G	0.6	0.0	-	5	1	.153
			B	0.6	0.0	-	-	-	.156
			L	6.0	0.6	4	-	-	.234
			G	1.4	0.5	-	3	1	.193
			B	1.4	0.4	-	-	-	.192
	SQD	SD2	L	6.0	0.9	8	-	-	.305
			G	1.4	0.5	-	3	1	.424
			B	1.4	0.4	-	-	-	**.428**
			L	0.6	0.2	2	-	-	.442
			G	0.6	0.0	-	5	1	.445
			B	0.6	0.0	-	-	-	.445
ASR	SD2	SQD	L	1.8	0.2	2	-	-	.111
			G	0.8	0.0	-	5	1	.089
			B	0.6	0.3	-	-	-	.109
			L	2.0	0.7	2	-	-	.134
			G	6.0	0.6	-	5	1	.129
			B	3.0	0.5	-	-	-	.129
	SQD	SD2	L	1.8	0.3	2	-	-	.248
			G	6.0	0.6	-	5	1	.217
			B	3.0	0.5	-	-	-	.242
			L	0.2	0.6	2	-	-	.275
			G	0.8	0.0	-	5	1	.264
			B	0.6	0.3	-	-	-	.266

5. Conclusions and future work

This paper presented a method for SCR that exploits the prosodic prominence of terms in the retrieval scoring process. Our method uses existing queries and relevance judgements to train a prosodic model whose predictions are subsequently incorporated into the Okapi BM25 function. Retrieval experiments over a collection of speech recordings in Japanese suggest that our prosodic retrieval method has the potential to provide significant improvements over standard text-retrieval methods, at least when manual transcripts of speech are used. In future work, we plan to test our method with a larger collection of speech recordings in English. We also plan to analyse the effects of ASR errors, speaking style variations, and the benefits of training prosodic models for individual speakers.

6. Acknowledgements

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.

7. References

[1] M. Larson and G. J. F. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 45, pp. 235–422, 2012.

[2] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Language and Cognitive Processes*, vol. 25, no. 7-9, pp. 905–945, 2010.

[3] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1, pp. 31–43, 2002.

[4] K. Spärck Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 1," *Information Processing & Management*, vol. 36, no. 6, pp. 779–808, 2000.

[5] G. Salton, "Mathematics and information retrieval," *Journal of Documentation*, vol. 35, no. 1, pp. 1–29, 1979.

[6] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, 1995, pp. 109–126.

[7] R. Silipo and F. Crestani, "Prosodic stress and topic detection in spoken sentences," in *Proceedings of the 7th International Symposium on String Processing and Information Retrieval*, ser. SPIRE'00, A Coruña, Spain, 2000, pp. 243–252.

[8] F. Crestani, "Towards the use of prosodic information for spoken document retrieval," in *Proceedings of the International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval*, ser. SIGIR'01. New Orleans, USA: ACM, 2001, pp. 420–421.

[9] B. Chen, H.-M. Wang, and L.-S. Lee, "Improved spoken document retrieval by exploring extra acoustic and linguistic cues," in *Proceedings Interspeech'01*, Aalborg, Denmark, 2001, pp. 299–302.

[10] C. Guinaudeau and J. Hirschberg, "Accounting for prosodic information to improve ASR-based topic tracking for TV broadcast news," in *Proceedings Interspeech'11*, Florence, Italy, 2011, pp. 1401–1404.

[11] D. N. Racca, M. Eskevich, and G. J. F. Jones, "DCU search runs at MediaEval 2014 Search and Hyperlinking," in *Proceedings of the MediaEval 2014 Multimedia Benchmark Workshop*, Barcelona, Spain, 2014.

[12] D. N. Racca and G. J. F. Jones, "DCU at the NTCIR-11 SpokenQuery&Doc task," in *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*. Tokyo, Japan: National Institute of Informatics, 2014, pp. 376–383.

[13] G. Lecorvé, G. Gravier, and P. Sébillot, "An unsupervised web-based topic language model adaptation method," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Las Vegas, Nevada, USA, 2008, pp. 5081–5084.

[14] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, ser. CIKM '04. New York, NY, USA: ACM, 2004, pp. 42–49.

[15] A. Rosenberg, "Modeling intensity contours and the interaction between pitch and intensity to improve automatic prosodic event detection and classification," in *Proceedings of Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 376–381.

[16] T. Mishra, V. K. R. Sridhar, and A. Conkie, "Word prominence detection using robust yet simple prosodic features," in *Proceedings Interspeech'12*, 2012.

[17] A. Rosenberg, "AuToBI - a tool for automatic ToBI annotation." in *Proceedings Interspeech'10*. Makuhari, Japan: ISCA, 2010, pp. 146–149.

[18] G. Christodoulides and M. Avanzi, "An evaluation of machine learning methods for prominence detection in french," in *Proceedings Interspeech'14*, 2014, pp. 116–119.

[19] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[20] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *20th International Conference on Pattern Recognition (ICPR)*. IEEE, 2010, pp. 3121–3124.

[21] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras, "Research directions in Terrier: a search engine for advanced retrieval on the web," *Novatica/UPGRADE Special Issue on Next Generation Web Search*, pp. 49–56, 2007.

[22] R. Song, J.-R. Wen, and W.-Y. Ma, "Viewing term proximity from a different perspective," Microsoft Research, Tech. Rep. MSR-TR-2005-69, May 2005.

[23] T. Akiba, H. Nishizaki, K. Aikawa, X. Hu, Y. Itoh, T. Kawahara, S. Nakagawa, and H. Nanjo, "Overview of the NTCIR-10 SpokenDoc-2 Task," in *Proceedings of the NTCIR-10 Workshop Meeting*, Tokyo, Japan, 2013, pp. 573–587.

[24] T. Akiba, H. Nishizaki, H. Nanjo, and G. J. F. Jones, "Overview of the NTCIR-11 SpokenQuery&Doc task," in *Proceedings of the NTCIR-11 Conference*, Tokyo, Japan, 2014, pp. 350–364.