Sentence Length Bias in TREC Novelty Track Judgements

Lorena Leal Bando & Falk Scholer
School of Computer Science
and Information Technology
RMIT University
Melbourne, Australia
{lorena.lealbando,falk.scholer}@rmit.edu.au

Andrew Turpin
Dept. of Computer Science
and Software Engineering
University of Melbourne
Melbourne, Australia
aturpin@unimelb.edu.au

ABSTRACT

The Cranfield methodology for comparing document ranking systems has also been applied recently to comparing sentence ranking methods, which are used as pre-processors for summary generation methods. In particular, the TREC Novelty track data has been used to assess whether one sentence ranking system is better than another. This paper demonstrates that there is a strong bias in the Novelty track data for relevant sentences to also be longer sentences. Thus, systems that simply choose the longest sentences will often appear to perform better in terms of identifying "relevant" sentences than systems that use other methods. We demonstrate, by example, how this can lead to misleading conclusions about the comparative effectiveness of sentence ranking systems. We then demonstrate that if the Novelty track data is split into subcollections based on sentence length, comparing systems on each of the subcollections leads to conclusions that avoid the bias.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Experimentation

Keywords

Sentence ranking, query-biased summaries, sentence length

1. INTRODUCTION

Automatic text summarisation is a well-studied field that includes a wide range of summary types, and methods to create summaries [11, 19, 21]. One method, extractive summarisation, constructs summaries by excerpting passages of documents that are deemed to be important, and then combines those passages to form the summary. As sentences typically express a single and complete idea, they are the dominant form of passage used in extractive summarisation. Sentences are particularly suitable as the building block of summaries for applications where the summary must be of a specified length. Thus, a key component of extractive summarisation methods is sentence ranking [7, 8, 17, 25, 30], where sentences are ranked according to some measure of their suitability for forming part of a summary, and the top m sentences are joined to create the summary.

Sentence ranking is a similar problem to document ranking, the cornerstone of the information retrieval (IR) discipline. In document ranking, documents in a collection are scored for their similarity to a posed query; for example, just as Google and Bing rank Web pages against a query. The IR field has a long history of rigorous methods for evaluating the effectiveness of document ranking systems, and so those well-established techniques can be adopted to assess sentence extraction methods that form the foundation of current summarisation methods.

The Cranfield methodology establishes a framework to gauge the effectiveness of IR systems in the context of document retrieval [5]. Such a framework involves: a text collection that is a set of documents to be searched; a set of topics which resemble user requests that could be answered using the documents in the collection; and judgements, given by external assessors, that estimate the "relevance" of a document with respect to each topic. Note that relevance is a complex concept and can vary from person to person, but in order to make experiments based on the Cranfield methodology tractable, relevance is typically simplified into a binary scale as determined by one person, and judgements are based only on the topical content or "aboutness" of a document. When comparing two systems, each topic is run by each system, a ranked list of documents is returned by each system, and then the relevance of documents in each list is scored and combined to give a quantitative measure of each system's effectiveness.

TREC 2002 introduced the Novelty track [10], which mainly aimed to investigate approaches to detect non-redundant information at the sentence level. The track supplied separate judgements on individual sentences for both "relevance" to the topics and "novelty". Therefore, these sentence level judgements can be used to evaluate sentence ranking systems that aim to rank based on relevance or novelty.

When assessing sentence ranking systems that are a preprocessor for a summarisation system, one could assume that the sentences should be sorted by "relevance" to a particular topic, and thus the Novelty track judgements can be

ADCS '12, December 5-6, 2012, Dunedin, New Zealand.
Copyright 2012 ACM 978-1-4503-1411-4/12/2012 ...$15.00.

used directly to score systems. However, depending on how the sentences are used to form a summary, relevance as determined by a single judge at a topic level may not be a valid basis for judging sentence ranking systems. In a Web context, where very short summaries called "snippets" are generated, sentences selected should indicate to a user whether to click through to the underlying document or not. Therefore, "indicativeness" may be a more suitable property on which to base judgements to compare sentence ranking systems. However, previous studies have used the Novelty track data to investigate passage retrieval [16] and summarisation approaches [18]; we therefore also assume that a sentence that is labelled as relevant for an assessor of the Novelty track is also indicative for assembling a summary. Thus, Novelty track relevance assessments are used for evaluating sentence ranking methods that form a pre-processor to snippet generation.

This paper exposes a weakness in using the Novelty track data in this way: namely there is a strong connection between sentence length and relevance. The next section examines the Novelty track data in detail, demonstrating this connection. Section 3 then shows an example of how using the typical Cranfield methodology and the Novelty track data demonstrates that the Vector Space Model adapted for sentences outperforms query-biased sentence selection. If a length component is introduced to each system; however, then the result is reversed and the query-biased system comes out on top. Note that both of these approaches are highly regarded in the literature, and not simply straw men. A second experiment in this section examines the effect of query expansion on the two systems, and the interplay of the length bias in assessing the superior system. Section 4 explains how the Novelty track data could be used to avoid making misleading system comparisons such as those in Section 3. In Section 5 we discuss our results, with conclusions and future work presented in Section 6.

2. NOVELTY TRACK DATA

The TREC Novelty track ran from 2002 to 2004 [10, 23, 24], and aimed to study passage retrieval to identify non-redundant content. In the first year of the track, assessors had specific constrains that led the identification of a very small proportion of relevant sentences. These restrictions included: to not select contiguous sentences; to not create topics; and to make judgements towards a short topic description. For these reasons organisers of the track suggested that the outcomes of the Novelty track 2002 should only be regarded as a pilot experiment [10]. Hence, we only use the data from 2003 and 2004 in our experiments.

The 2003 and 2004 Novelty tracks are more homogeneous in terms of constructing topics and gathering judgements [23, 24]. NIST assessors created 50 topics from the AQUAINT newswire collection for each year of the track, and identified relevant documents for the topic by employing the WebPRISE information retrieval system. Novelty 2003 was comprised of 25 relevant documents per topic, while Novelty 2004 tallied more than 25 documents for some topics, as irrelevant documents were intentionally included. Irrelevant documents were added in order to increase the complexity of the task for participants, relative to 2003.

Documents in the AQUAINT collection contain an id corresponding to the date of authorship. Relevant documents were sorted according to this identifier, and split into sentences. All document sentences were pooled into a single document for judgement, so an assessor inspected documents chronologically instead of their ranked position provided by the retrieval system. Assessors were asked to distinguish relevant sentences in a topic, and to identify those that were novel. That is, relevance and novelty judgements were made separately. Assessors were able to select any number of relevant and novel sentences per topic or document.

Despite the fact that the main goal of the Novelty track was to study techniques to avoid redundant content, the availability of relevance judgements at the sentence level makes the track appealing for other types of applications. For instance, previous research has employed the Novelty track data to evaluate machine learning approaches for snippet generation [18] and passage retrieval tasks assisted by statistical query expansion [16]. The former work found that the selection of features among Novelty data sets from 2002 to 2004 were not robust. Losada [16], on the other hand, found that query expansion effectively assisted the identification of relevant sentences. However, his findings are applicable for sentences within a set of documents regarding the same topic, rather than individual documents.

2.1 Length Bias

Previous research has shown that the relevance of a document tends to increase with its length, since long documents may include information related to not only one but several requests. Consequently, users are prone to select these documents as relevant. The document length bias was investigated in early TREC conferences to determine the effectiveness of a retrieval system [22]. It was found that if a system took into account the length of a document in the ranking process, it could outperform those that did not. This length bias has not previously been investigated for sentence ranking tasks.

Approaches to extractive summarisation rely on the occurrence of significant terms, cue words, query words, or title words of documents for ranking sentences (discussed in more detail in Section 3.1). Therefore, long sentences are more likely to contain these terms in comparison to short sentences, and thus to be scored more highly with respect to being included in summaries.

To avoid ranking long sentences ahead of short sentences purely because of length, sentence scores can be normalised by dividing by the total number of words in each sentence [26]. Normalisation emerged as a mechanism to minimise the effect of retrieving long documents [22] for the document ranking problem. Another approach is to ignore brief sentences, since they might not include relevant content due to their length [13].

2.2 Length Bias in the Novelty Track

A simple approach to detect a length bias in the data set is to count the number of words in relevant and irrelevant sentences. It should be noted that assessors did not provide relevance judgements for sentences in irrelevant documents, which were intentionally included in the Novelty 2004. Thus, there is a potential pool of relevant sentences that were not judged as such in the collection. To avoid the confounds that these might introduce in our analysis, we discarded all documents (hence sentences within those documents) that do not contain at least one relevant sentence. This ensures that every sentence in our analysis of the Novelty track 2004

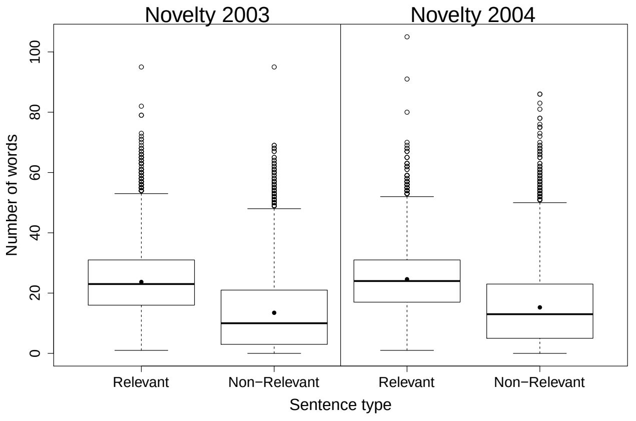

Figure 1: Distribution of the number of words in each sentence (including stopwords) in the Novelty track data sets (2003 and 2004). Boxes are 25th and 75th percentiles, bar is median, whiskers and circles show extremities, and the dot represents the mean.

Table 1: Composition of the two Novelty track collections from 2003 and 2004 for known judged documents. Sentence length is the mean number of words (including stopwords.)

	Novelty track	
Feature	2003	2004
Number of relevant sentences	15,557	8,343
Number of non-relevant sentences	24,263	28,628
Length of relevant sentences	23.69	24.59
Length of non-relevant sentences	13.47	15.26
Number of topics	50	50
Number of documents	1,250	1,070

data has been judged.

Figure 1 shows the sentence lengths on the two collections. The mean length of non-relevant sentences is significantly less than the mean length of relevant sentences (*t-test* $p < 0.001$), with details given in Table 1. There is a clear connection between length and relevance in the 2003 and 2004 Novelty track data. The next section examines how this length bias manifests in sentence ranking methods.

3. SYSTEM COMPARISONS

This section describes two systems for ranking sentences, and compares them using the Cranfield methodology on the Novelty track 2003 and 2004 data ignoring possible sentence length bias. We then introduce a length bias component into each system and study the effects.

3.1 Extractive and Query-biased Summarisation

An extractive summarisation method ranks sentences based on evidence collected either directly from documents or from shallow sentence attributes. The former mechanism involves word frequency statistics to determine significant words from the document [17], to identify title and heading words or cue words [7]. Sentences are scored according to the occurrence of such terms. Early summarisation approaches have employed evidence from documents. However, modern text collections may have metadata information [12] and anchor text [2] available, which can be used for assisting Web page summarisation.

Shallow sentence attributes, on the other hand, are independent from the document vocabulary and concentrate on superficial features such as the position [4] or length of sentences [13], and word formatting [30]. In order to take advantage of both document content and shallow sentence features, these can be merged using a linear combination where constants tune the value that each approach contributes to the final score of a sentence.

Query-biased summarisation (also called query-dependent, query-specific or query-relevant) is a type of extractive summarisation that favours the selection of sentences, or passages, that contain query terms. In large text collections, these summaries are helpful for guiding users to dismiss irrelevant documents and to inspect those that appear likely to be relevant given the summary [25]. Query-biased summarisation can rely on simple heuristics that count for query terms occurrence, or document ranking functions adapted for sentence selection. The following subsections explain both approaches in detail.

Query Term Occurrence

Tombros and Sanderson [25] introduced a score that depended on the appearance of query terms for the ranking of candidate sentences for summaries of Web search results. This score was computed in a similar fashion as the clusters of significant words proposed by Luhn [17]. The query-biased score for a sentence s is calculated as:

$$\mathrm{QB}_s = \frac{(|qt|)^2}{|q|} \tag{1}$$

where $|qt|$ is the number of unique query terms in sentence s and $|q|$ is total number of words in the query. Other variants for employing query terms occurrence include the count of repeated query terms, and the longest contiguous sequence of query terms in a sentence [27]. In commercial applications, query counting occurrence is used for extracting "query-relevant" parts from documents. These excerpted parts are not only useful for displaying snippets of search engine results, but also for detecting duplicate documents without the need of analysing the whole document [9].

VSM

A variety of retrieval models have been proposed in the literature such as the Vector Space Model (VSM), the Okapi BM25 similarity function, and Language Models [6]. By treating each sentence as a "document" is straightforward to apply these to score sentences relative to a query [1, 8, 16, 28]. For example, the cosine similarity function in the VSM calculates the Euclidean distance between weighted document and query vectors. That is, the shorter the distance between both vectors, the more similar a query is to a document (or to a sentence). Allan et al. [1] adapted the VSM for determining the similarity of a sentence to a query as follows:

$$R(s|q) = \sum_{t \in q} \log(f_{(t,q)} + 1) \log(f_{(t,s)} + 1) \log\left(\frac{n+1}{0.5 + f_t}\right) \tag{2}$$

where $f_{(t,q)}$ and $f_{(t,s)}$ are the occurrence of term t in query q and sentence s, respectively. The number of sentences in the collection is given by n, and f_t is the number of sentences in which the term t appears.

Similar applications of other retrieval models for sentence selection have been studied. Nevertheless, comparisons in previous work have not shown significant differences between the effectiveness of VSM compared to the Okapi BM25 similarity function [16], or Language Models with the Kullback-Leibler divergence [1, 14, 16].

The generation of query-biased summaries is not restricted to statistical methods; machine learning approaches can also be used [18, 29]. However, the focus of this paper is not an exhaustive comparison of query-biased summarisation methods, so we restrict our attention to two widely-used approaches: the query-biased score (QB) and the VSM model, described above.

3.2 Baseline Systems

The QB approach consists of counting occurrences of query terms in a sentence as defined in Equation 1. The VSM approach employs the vector space model adaptation for sentence retrieval as applied by Allan et al. [1]. Since terms are generally not repeated in typical Web search queries, the VSM is simplified as follows:

$$R(s|q) = \sum_{t \in q} \log(f_{(t,s)} + 1) \log\left(\frac{n+1}{0.5 + f_t}\right) \qquad (3)$$

As we are addressing a single-document summarisation problem, the parameter n in this equation is the number of sentences in a given document, instead of the number of sentences in a collection as used by Allan et al. [1].

Both the QB and VSM approaches allow the scoring of all sentences in a document, and the top m sentences are selected for inclusion in a summary. Ties in scores can be broken by choosing sentences that occur closer to the beginning of the document. These two baseline ranking approaches are identified as QB-Pos and VSM-Pos. A second simple way to resolve ties is in the favour of longer sentences. These approaches are labelled QB-Len and VSM-Len.

3.3 Results

The Novelty 2003 and 2004 track supplied relevance judgements at the sentence level, thus a simple way to measure the performance of ranking methods is to calculate the proportion of returned sentences that are relevant in a document. Similar to document ranking evaluation approaches, we adopt P@m as the measure to quantify the performance of summarisation methods, where m is the number of returned sentences. In our experiments we use m=2, as the aim is to assemble short excerpts. Thus, for any topic we can average the P@2 for each document for that topic, and compare the means between different sentence ranking schema. From the Novelty data set, we included all documents that have at least m relevant and m non-relevant sentences as part of our test collection.

The top two rows of Table 2 show the results of the two baseline methods using the "title" field of the TREC Novelty topics as a query. The title averages a length of three words for both Novelty 2003 and 2004, similar to current Web queries [3]. While VSM-Pos outperformed QB-Pos significantly for Novelty 2003 (t-test, $p < 0.001$), the percentage change for Novelty 2004 is not significant ($p > 0.05$).

Rows four and five show the two baseline methods, but here tied sentence-ranking scores are resolved based on decreasing sentence length. Note that the P@2 values are higher (between 7% and 26%), which is to be expected as now longer sentences are being favoured, and we know that, in general, longer sentences are more likely to be relevant in these collections. All increases are statistically significant (*t-test*, $p < 0.001$). Not only are all four numbers higher than when not using length to break ties, but now the QB based system is better than the VSM system (*t-test*, $p = 0.002$ and $p = 0.006$ for Novelty 2003 and 2004, respectively). That is, taking length into account has reversed the result of the original experiment.

Table 2: P@2 results of the four methods with original and expanded queries. The method Len ignores the query and consider longer sentences, and a method that randomly selects two sentences (Random). An $**$ indicates statistical significance (paired t-test) of $p < 0.001$ and $*$ of $p < 0.01$.

Method	Novelty 2003	Novelty 2004
Original query *(title)*		
QB-Pos	0.61	0.52
VSM-Pos	0.68	0.53
Change	10%**	2%
QB-Len	0.77	0.58
VSM-Len	0.75	0.56
Change	-3%*	-3%*
Only length		
Len	0.72	0.52
Random	0.44	0.27
Expanded query *(title and narrative)*		
QB-Pos	0.73	0.60
VSM-Pos	0.75	0.62
Change	1%	3%
QB-Len	0.79	0.63
VSM-Len	0.75	0.62
Change	-4%**	-1%

As a point of comparison, we ranked sentences using only their length and ignore any score that involves the query. This simple baseline is called Len, row seven in the above table. It can be noted that for the Novelty 2003 VSM-Pos, which outperformed QB-Pos, performed more poorly than Len. This confirms a strong length bias in the relevance assessments in this data set as suggested by Metzler and Kanungo [18] when using machine learning approaches on the Novelty 2003 data. In the case of the Novelty 2004, the length effect is more moderate compared to QB-Pos and VSM-Pos.

Query Expansion

It has been suggested that automatically expanding a query can assist in the selection of sentences for passage retrieval tasks [16]. We were also investigating the use of query expansion to improve sentence ranking for snippet generation when we came across the length bias in the TREC Novelty track 2003 and 2004 data described in this paper. While there are some sophisticated query expansion methods we could apply to QB-Pos and VSM-Pos, in this paper we add the "narrative" field of the TREC Novelty topics to the "title" field as ready-made expanded queries. The narrative field of a TREC Novelty topic consists of a verbose statement of the information need that corresponds to the short title query; using this as a proxy for query expansion is therefore equivalent to considering a case where a relevance feedback system has elicited an extended description of an information need from a user.

The bottom section of Table 2 shows the results for these queries. In all cases, the systems using the expanded queries scored higher than their non-expanded counterparts. And again, ignoring length puts VSM ahead of QB as the method of choice, while adding length reverses the result. Row eight in the same table shows the performance of randomly selecting sentences (Random). This method significantly achieved poorly in comparison to the Len approach and both VSM and QB using original or expanded queries ($p < 0.001$).

4. ISOLATING SENTENCE LENGTH BIAS

While it has been proposed that short sentences could simply be ignored in the construction of summaries [13], for query-biased summaries in space-limited environments such as search result pages, short sentences can be valuable. We therefore investigate how to isolate the sentence length bias factor for evaluation purposes. Our approach is similar to that used by Singhal et al. [22] for document retrieval. In their approach, documents were grouped by their length (measured in bytes). In contrast, we bucketed sentences of specific lengths measured in words to attempt to counter the effects of a length predisposition.

We obtained the average length (μ) of relevant and irrelevant sentences in the Novelty data, as well as the standard deviation (σ). This is $\mu = 17$ and $\sigma = 12$ words for both Novelty track 2003 and 2004 (according to sentence statistics listed in Table 1). Based on this information we classified sentences in three buckets: l_1, where sentences have between 5 ($\mu - \sigma$) and 13 words; l_2 contains sentences from 14 to 20 words; and l_3, containing sentences between 21 to 29 ($\mu + \sigma$) words. Given that each bucket contains sentences of different lengths, the amount of information for each sentence may vary from bucket to bucket. A summary composed of two sentences from bucket l_3 is longer – and potentially more indicative – than a summary composed of two sentences from bucket l_1. To account for this, we adapt the value of m in P@m for each bucket. Specifically, we used P@4, P@3 and P@2 for measuring effectiveness in each of three buckets l_1, l_2 and l_3, respectively.

Table 3 lists the number of documents that exist in each bucket. This can be seen as splitting the Novelty data set into several subcollections, where a document has at least m relevant sentences of length l_i. For comparison purposes, for each bucket we collected the m longest sentences (Len approach), and randomly selected m sentences of length l_i (Random approach).

The results of evaluating the different summarisation approaches using the length-bucketing approach are shown in Table 4. Contrary to the results in Section 3.3, we did not find significant differences between QB-Pos and VSM-Pos when employing either the original query (rows 1 and 2), or its expanded version (rows 4 and 5). By analysing buckets of length l_2 and l_3 in the Novelty track 2003 data, the expansion did not reveal any improvement over the original query when the sentence length feature is isolated. However, for the same buckets in the Novelty track 2004 the differences were significant (t-test, $p < 0.05$). For short sentences (bucket l_1), the difference when using query expansion is significant ($p < 0.05$) for both the Novelty 2003 and 2004 data, with percentage changes from 4% to 8%.

Table 3: Number of documents in each bucket to compute P@m.

Bucket	m	Novelty 2003	Novelty 2004	Total documents
l_1	4	629	588	1217
l_2	3	722	682	1404
l_3	2	1020	983	2003

Given that QB-Pos is not significantly different from VSM-Pos, we use the former approach for comparison against Len. For buckets of length l_1, Len performs significantly better than using the expanded query in both data sets ($p < 0.05$). For the remaining buckets, we noted that QB-Pos with query expansion performs significantly better than the Len method ($p < 0.001$). Finally, we observed that the Len method against the Random differs performance using buckets of length l_1 in the Novelty track 2003 and 2004 ($p < 0.001$). For the other buckets both methods achieved similar P@m scores ($p > 0.05$).

5. DISCUSSION

We have demonstrated that there is a length bias in the TREC Novelty track data, where relevant sentences tend to be longer sentences. As a demonstration of how this bias could lead to misleading claims of system effectiveness, we report novel system comparisons on two leading methods for the generation of query-biased summarisation. The thesis of this paper is not to promote one or other of the methods studied, but rather to indicate a potential weakness in using the Novelty track sentence relevance judgements to assess systems that rank sentences.

It can be observed that, in general, a Random approach did not outperform the QB- or VSM-based methods. The exceptional case when this occurs is in the Novelty 2003 data for buckets of length l_1. This suggests that for short sentences (5-13 words) in this data set, other constrains should be applied to more accurately discern relevant sentences. The values to define thresholds in buckets were gathered from sentences in the Novelty track as outlined in Section 2.2. Given that a summary is restricted in size, we assumed that 4, 3 and 2 sentences were representative for their corresponding buckets. However, we did not explore other parameter settings for the optimal number of words to be considered in each bucket.

In this paper we have focussed solely on evaluating the sentence ranking problem using TREC Novelty track data. The Text Analysis Conference (TAC), formerly known as the Document Understanding Conference (DUC), has investigated different summarisation styles and proposed several intrinsic evaluation methodologies since 2001. Such methodologies assessed automatic summaries in terms of vocabulary

Table 4: P@m values for buckets of sentences of length l.

		Novelty 2003			Novelty 2004		
		P@2	P@3	P@4	P@2	P@3	P@4
		l_3	l_2	l_1	l_3	l_2	l_1
Original	QB-Pos	0.709	0.613	0.306	0.489	0.383	0.188
(*title*)	VSM-Pos	0.705	0.614	0.308	0.492	0.384	0.187
	Change	-0.6%	0.2%	0.6%	0.5%	0.2%	-1.0%
Expanded	QB-Pos	0.708	0.616	0.327	0.520	0.401	0.196
(*title and*	VSM-Pos	0.699	0.613	0.333	0.512	0.404	0.198
narrative)	Change	-1.3%	-0.4%	1.7%	-1.4%	0.6%	1.5%
Len	—	0.643	0.572	0.398	0.412	0.338	0.209
Random	—	0.636	0.562	0.335	0.403	0.327	0.175

overlap [15], or content matching units [20], against a set of ideal summaries. This set is comprised of abstracts authored by assessors who may merge ideas into a single sentence or paraphrase content, for example. Hence, the framework provided by TAC/DUC cannot be used straightforwardly to evaluate sentence ranking methods. Despite the fact that the Novelty track offers relevance judgements of sentences, further research is required to evaluation on such assessments reliable, particularly for applications where the use of these judgements deviate from *ad-hoc* and novelty tasks, which were the main aims of the track.

6. CONCLUSIONS AND FUTURE WORK

In this paper we investigated the way in which the length of a sentence affects the selection of relevant sentences, specifically in the Novelty track data, which has been used in the IR field to evaluate sentence ranking and snippet generation. This fact calls into question past conclusions on the effectiveness of sentence ranking approaches for tasks such as summarisation or passage retrieval that were based on this data set.

We found that using a short baseline query, similar to current Web queries, ranking methods performed significantly better when employing the sentence length component. Simulating a simple query expansion approach by using the narrative field of Novelty track topics showed that significant improvements from 7% to 26% can be achieved when ignoring the impact of sentence length. However, this advantage disappears when sentence length is included as a component in the ranking method.

We proposed an alternative method for evaluating sentence ranking methods for the construction of query-biased summaries by measuring P@m of sentences of similar length. This approach avoids any length predisposition of sentences. Under this controlled evaluation framework, both original and expanded query-biased approaches were shown to outperform a baseline where sentences were selected only based on length. Similarly, both original and expanded queries were in general able to outperform a random selection of sentences, except for the 2003 Novelty data for short sentences (bucket l_1). We plan to study other formal approaches of query expansion in future work.

Furthermore, our analysis demonstrated that relevance judgements from the Novelty track might not be entirely suitable for evaluating summarisation approaches. These assessments were designed for *ad-hoc* tasks and it is not clear whether they can straightforwardly be treated as surrogates of indicative content. For example, snippets that are displayed by search engines need to be concise, so that users can make a decision to click on a result or to keep reading the list of results. In a follow-up study, we plan to directly investigate the indicative value of relevant sentences in the Novelty track for the assembling of query-biased summaries, and the relation between sentence length and indicative content.

7. ACKNOWLEDGEMENTS

This research is supported in part by ARC Grant FT0991326 (AT) and CONACyT scholarship (201099).

8. REFERENCES

[1] J. Allan, C. Wade, and A. Bolivar. Retrieval and Novelty Detection at the Sentence Level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 314-321, ACM Press, 2003.

[2] E. Amitay and C. Paris. Automatically Summarising Web Sites: is there a way around it? In *Proceedings of the ninth international conference on Information and knowledge management*, 173-179, ACM Press 2000.

[3] M. Bendersky and W. Bruce Croft. Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the 2009 workshop on Web Search Click Data*, WSCD '09, 8-14, ACM Press, 2009.

[4] R. Brandow, K. Mitze, and L. F. Rau. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing & Management*, 31(5):675-685, 1995.

[5] C. Cleverdon. The Cranfield Tests on Index Language Devices. *ASLIB Proceedings*, 19:173-194, 1967.

[6] W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information retrieval in practice.* Addison Wesley, 2009.

[7] H. P. Edmundson. New Methods in Automatic Extracting. *Journal of the ACM*, 16(2):264-285, 1969.

[8] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing Text Documents: Sentence Selection and Evaluation Metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 121-128. ACM Press, 1999.

[9] B. Gomes and B. T. Smith. Detecting Query-specific Duplicate Documents, Patent No. 6,615,209 B1, 2003.

[10] D. Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of TREC 2002*, 2002.

[11] K. Spärck Jones. Automatic Summarising: The state of the art. *Information Processing & Management*, 43(6):1449-1481, 2007.

[12] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving Search Results Quality by Customizing Summary Lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 701-709. ACL, 2008.

[13] J. Kupiec, J. Pedersen, and F. Chen. A Trainable Document Summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, 68-73. ACM Press, 1995.

[14] X. Li and W. B. Croft. Novelty Detection Based on Sentence Level Patterns. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 744-751, ACM Press, 2005.

[15] C. Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the workshop on text summarization*, 74-81, 2004.

[16] D. E. Losada. Statistical Query Expansion for Sentence Retrieval and its Effects on Weak and Strong Queries. *Information Retrieval*, 13(5):485-506, 2010.

[17] H. P. Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2(2):159-165, 1958.

[18] D. Metzler and T. Kanungo. Machine Learned Sentence Selection Strategies for Query-biased Summarization. In *Proceedings of SIGIR Workshop on Learning to Rank for Information Retrieval*, 40-47, 2008.

[19] A. Nenkova and K. McKeown. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103-233, 2011.

[20] A. Nenkova and R. Passonneau. Evaluationg Content Selection in Summarization: The Pyramid Method. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 145-152, 2004.

[21] C. D. Paice. Constructing Literature Abstracts by Computer: Techniques and Prospects. *Information Processing & Management*, 26(1):171-186, 1990.

[22] A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document Length Normalization. *Information Processing & Management*, 32:619-633, 1996.

[23] I. Soboroff. Overview of the TREC 2004 Novelty Track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

[24] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, 2003.

[25] A. Tombros and M. Sanderson. Advantages of Query biased Summaries in Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 2-10. ACM, 1998.

[26] Y. Tsegay, S. Puglisi, A. Turpin, and J. Zobel. Document Compaction for Efficient Query Biased Snippet Generation. In Mohand Boughanem, Catherine Berrut, Josiane Mothe, and Chantal Soule-Dupuy, editors, *Advances in Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, 509-520. Springer Berlin / Heidelberg, 2009.

[27] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast Generation of Result Snippets in Web Search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 127-134. ACM Press, 2007.

[28] R. Varadarajan and V. Hristidis. A System for Query-specific Document Summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, 622-631, ACM Press, 2006.

[29] C. Wang, F. Jing, L. Zhang, and H. Zhang. Learning Query-biased Web Page Summarization. In *Proceedings of the 16th ACM conference on Conference on information and knowledge management*, 555-562. ACM Press, 2007.

[30] R. W. White, J. M. Jose, and I. Ruthven. A Task-oriented Study on the Influencing Effects of Query-biased Summarisation in Web Searching. *Information Processing & Management*, 39(5):707-733, 2003.