Position-Based Contextualization for Passage Retrieval

David Carmel*
Yahoo! Lab
Haifa, 31905, Israel
david.carmel@ymail.com

Anna Shtok†
Industrial Engineering and Management
Technion, Haifa 32000, Israel
anabel@tx.ac.il

Oren Kurland
Industrial Engineering and Management
Technion, Haifa 32000, Israel
kurland@ie.technion.ac.il

ABSTRACT

We present a novel *contextualization* approach for passage retrieval. The core principle is to let any occurrence of a query term in a document affect the passage retrieval score, whether the occurrence is in the passage or not. This effect is controlled by the distance between the term occurrence and the passage. Empirical evaluation demonstrates the merits of our approach; the resultant retrieval performance substantially transcends that of previously proposed passage retrieval methods, including those that use various contextualization approaches.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Retrieval models

General Terms: Algorithms, Experimentation

Keywords: Passage Retrieval, Contextualization, Term Proximity

1. INTRODUCTION

Traditional ad hoc retrieval methods deal with the task of finding the documents that are most relevant to the user query. However, documents might be long and can cover many topics. In such cases, retrieving the most specific query-pertaining text units in the document, i.e. *passages*, rather than the document as a whole, can be of much merit [17, 10, 3]. Passage retrieval also serves as an intermediate phase in several other tasks such as question answering [18], entity oriented search [6] and text summarization [13].

There is an important difference between relevance estimation for documents and for passages. Document relevance can often be effectively estimated independently of other documents in the collection. On the other hand, passages are shorter units of text. Hence, effective passage-relevance estimation calls for *contextualization*; that is, the consideration of the passage context. Several types of contextualization, typically used for passage retrieval, are the text unit containing the passage, the document the passage belongs to, and reference documents [4, 14, 9, 15].

We present a novel contextualization approach for passage retrieval. The approach leverages the fundamental principle underlying the *positional language model (PLM)* [12] that was introduced for the document retrieval task. The key idea behind PLM is that any term in a document is represented as a density function which expresses the probability of finding the term at any position in the document. Similarly, our model estimates the probability of finding the query terms within the passage while considering all term occurrences in the document. Yet, our approach serves for passage retrieval rather than for document retrieval. The method we devise lets any occurrence of a query term in the document affect the passage score regardless of whether the occurrence is actually in the passage. This effect depends on the distance between the query term occurrences and the passage. Thus, the whole document, or more precisely, all query term occurrences in the document, provide context for the passage.

Evaluation performed with the INEX focused-retrieval benchmarks shows that our approach substantially improves over previously proposed passage retrieval methods, including those that use various contextualization techniques.

2. RELATED WORK

Several contextualization approaches for passage retrieval were examined in the past. A commonly used context for passages is their containing document [1, 14, 9]; e.g., using term counts in the document for "smoothing" those in the passage. In contrast to our approach, distances between the query terms and the passage at hand are not considered. Furthermore, we show that our method substantially outperforms this approach.

Another type of passage contextualization is considering its neighbor passages in the document [9]. Our approach, which lets query terms in the entire document affect the retrieval score of any passage, is shown to outperform this contextualization method.

The structure of an XML document [4, 15] and hyperlinks [15] were also used for passage contextualization. Such approaches are complementary to ours that considers term proximity in unstructured text.

The work most related to ours is that of Beigbeder who applied proximity scoring for focused retrieval [5]. Each position in the text is assigned with a proximity value depending

[*] This work was done while David Carmel was at IBM Haifa Research Lab.

[†] This work was done while Anna Shtok interned at IBM Haifa Research Lab.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright 2013 ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507865 .

on its distance from the query terms. These values can be summed on any range of text, and therefore can be applied for passage retrieval. The proximity value of a position to a term is determined based on the nearest occurrence of the term to the position, and the proximity value to the query is an aggregation of the position proximity scores to all query terms using fuzzy logic rules. In contrast, to the best of our knowledge, our work is the first to use *all* occurrences of the query terms in a document, and their distances from the passage at hand, for passage retrieval contextualization.

3. PASSAGE RETRIEVAL MODELS

3.1 Baseline approaches

Let q and d denote a query and a document respectively. Let $p \overset{def}{=} (p.s, p.e)$ be a passage which spans from position $p.s$ to position $p.e$ in the document. A commonly used passage scoring model, referred to here as **psg**, is based on the well known tf-idf-based ranking approach that was found to be highly effective in the passage retrieval domain, compared to several other ranking alternatives [11, 2]:

$$Score_{psg}(p;q) \overset{def}{=} \sum_{t \in q \cap p} tf(t,p) * idf(t); \tag{1}$$

$tf(t,p) \overset{def}{=} \log(\#occ(t,p)+1)$, where $\#occ(t,p)$ is the number of occurrences of term t in p; $idf(t) = \log(\frac{N}{N_t})$; N is the total number of documents in the collection, and N_t is the number of documents containing t.

Since passages are short units of text, the potential for vocabulary mismatch between a query and a relevant passage is quite high. Thus, for estimating passage relevance it is beneficial to use information from the passage *context* in addition to that in the passage itself. Using the document containing the passage is a commonly used contextualization approach [7, 1, 14]. Specifically, the (normalized) passage score assigned by Equation 1 to passage p is "smoothed" with the (normalized) retrieval score of the document d to which p belongs. Formally, the **psgDoc** method scores p by:

$$Score_{psgDoc}(p;q) \overset{def}{=} (1-\lambda)\frac{Score_{psg}(p;q)}{\sum_{p' \in d_p} Score_{psg}(p';q)} + \lambda\frac{Score_{doc}(d_p;q)}{\sum_{d' \in D_n} Score_{doc}(d';q)}; \tag{2}$$

λ is the contextualization parameter; d_p is the document containing p; $Score_{doc}(d;q)$ is document d's retrieval score which in our experiments is assigned by the state-of-the-art OKAPI-BM25 [16] method; D_n is the set of top-n scored documents in the corpus.

The passage context can be further refined by using its surrounding passages in the document [9]. For example, the passage score can be further smoothed with that of its neighbor passages, yielding the **psgNeighbor** method:

$$Score_{psgNeighbor}(p;q) \overset{def}{=} (1-\lambda_l-\lambda_r)Score_{psgDoc}(p;q) + \lambda_l Score_{psgDoc}(p_l;q) + \lambda_r Score_{psgDoc}(p_r;q), \tag{3}$$

where p_l and p_r are the two neighbor passages of p, from left and right, respectively; λ_l and λ_r are the contextualization parameters. This method can be generalized so as to consider the scores of the $k > 1$ neighbor passages from left and right with the values of the corresponding contextualization parameters decreasing with increasing distance from p.

3.2 A novel contextualization approach

We consider a novel contextualization approach for passage retrieval that is based on leveraging a fundamental principle underlying the locality-based similarity [8], and its successor positional language model (PLM) [12] that were proposed for document retrieval. PLM defines for each term position x in document d the probability $Pr(t|d,x)$ that term t will be generated in this position. For an occurrence of t in position o in d, $f_t(o,x)$ is the generation probability of t "propagated" from o to x. All occurrences of t in d affect the generation probability of t in x based on their distance from x. The more occurrences of t in d, and the closer these occurrences to x, the higher the generation probability of t at x.

For typical propagation functions, often called kernels, $f_t(o,x)$ is estimated based on the inverse of the distance of x from o. A commonly used kernel function is the Gaussian,

$$f_t^G(o,x) \overset{def}{=} e^{-\frac{(o-x)^2}{2\sigma^2}},$$

where σ is a free parameter. Note that the value decays with increasing distance from o. This kernel was shown to work well when applied for document retrieval [8, 12].

Another kernel we experimented with is the Trapezoid. The assigned value is 1 if x is in the same passage containing the position o, and otherwise decays linearly with increasing distance from the passage borders. Formally, let p^o be the passage containing the term position o. Then,

$$f_t^T(o,x) = \begin{cases} 1 & p^o.s \le x \le p^o.e \\ \max(0, 1-\frac{1}{\sigma}*(p^o.s-x)) & x < p^o.s \\ \max(0, 1-\frac{1}{\sigma}*(x-p^o.e)) & x > p^o.e. \end{cases}$$

For the passage retrieval task that we address, we use the underlying principle of PLM, described above, to estimate the contribution of the occurrences of all query terms in the document to the passage score. Given an occurrence of query term t in position o, we define its contribution to the score of passage p by computing the area under the kernel function between the passage borders. This area can be thought of as reflecting the probability of "generating" t in the given passage. The score of passage p for query q is then defined to be the sum of the contributions to p's score of all query-terms occurrences in the document. This is our **PLM** method:

$$Score_{PLM}(p;q) \overset{def}{=} \sum_{t \in q \cap d} idf(t) * \left[\sum_{o \in Occ(t,d)} \int_{p.s}^{p.e} f_t(o,x)dx\right]; \tag{4}$$

$Occ(t,d)$ is the set of all occurrences of t in d.

For both kernel functions described above, the steepness parameter σ controls the amount of contextualization. When the value of the steepness parameter is approaching zero, contextualization is reduced as only term occurrences in the passage, or that are very close to the passage, affect its score, while query term occurrences that are far from the passage do not contribute to its score. In contrast, when the value of the steepness parameter grows to infinity, the PLM method scores all passages equally (modulo passage length normalization, see details below), as any query term occurrence in the document contributes to all passages equally.

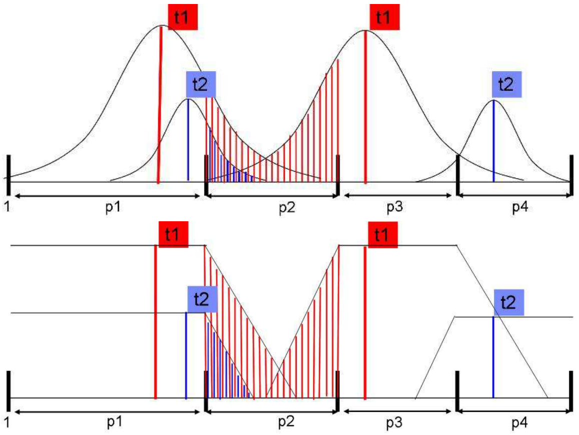

Figure 1: Example: Contextualized PLM passage scoring with Gaussian kernels (top) and Trapezoid kernels (bottom). Kernel heights correspond to terms' importance (*idf*). Paragraph $p2$ does not contain any of the query terms. However, its score is affected by the two neighbor occurrences of t_1 and the left neighbor occurrence of t_2.

The PLM method incurs bias in favor of long passages, as they span over large parts of the document. To ameliorate this bias, passage scores should be normalized. The normalization approach we applied is to approximate the area under the kernel curve by dividing the passage to k equal length intervals, and summing the kernel values in the $k+1$ interval borders; k is fixed and independent of the passage length hence the contribution of a kernel function to the passage score is independent of the passage length.

Figure 1 demonstrates the PLM score computation. The scheme illustrates the case of a two-terms query (t_1, t_2), and a document with 4 non-overlapping paragraphs. Paragraph $p2$ does not contain any of the query terms; however, its score is affected by the two neighbor occurrences of t_1 and by the left neighbor occurrence of t_2.

Finally, as is the case for the baseline scoring methods from Section 3.1, the (normalized) PLM score is further smoothed with the (normalized) document score to yield the overall passage score used for the final ranking.

4. EVALUATION

4.1 Experimental setup

Dataset. To evaluate our retrieval model, we conducted a set of experiments using the INEX dataset for the *focused tasks* of 2009 and 2010 [10, 3]. The INEX focused task requires systems to find the most focused results (i.e., shortest passages) that satisfy an information need, without returning overlapping passages. The dataset contains $2,666,190$ English Wikipedia articles, converted to XML format, with 68 judged topics for the 2009 task, and 52 judged topics for the 2010 task. Articles were judged by their relevance to the topics, and relevant parts of the articles were labeled explicitly. A quantification of the character-based overlap between retrieved passages and the labeled parts of the articles is used for system evaluation. Precision is measured as the portion of retrieved text that was labeled and recall is measured as the portion of all labeled text that has been retrieved. The interpolated precision measure is the precision score at a selected recall level x (iP[x]) and the mean average interpolated precision (MAiP) is the mean over 101 standard recall points. The official focused task measure for system evaluation, in 2009, was the interpolated precision at 1% recall level (iP[.01]) [10]. Statistical significance of performance differences is determined using the two tailed paired t-test at a confidence level of 95%.

Passage Scoring Methods. We employ a standard passage-based retrieval approach [10]. First, we create an initial document set from the 1500 documents that are the most highly ranked for a topic by using the BM25 retrieval method with default parameter settings ($b = 0.2, k1 = 0.6$); titles of topics serve for queries. Then, we use the paragraphs in these documents as retrievable passages[1]. These passages are ranked by one of the methods specified below. The 1000 top ranked passages constitute the final result list.

We experimented with several methods for passage scoring:

- psg: The standard tf.idf-based method specified in Equation 1.
- psgDoc: The method that uses the document retrieval score for smoothing the passage retrieval scores. (See Equation 2.) We tuned the interpolation parameter (with respect to values in the range $\{0, 0.1, \ldots, 1\}$), and report the performance values for $\lambda = 0.9$, which results in optimal performance for both benchmarks.
- psgNeighbor: The method that smooths the score assigned to the passage by the psgDoc method ($\lambda = 0.9$) with the scores assigned by psgDoc to its neighbor passages. (Refer to Equation 3.) We tuned the contextualization parameters (with respect to values in the range $\{0, 0.25, \ldots, 1\}$), and report the performance for $\lambda_l = \lambda_r = 0.25$, a setting that results in improved performance with respect to other settings for the two benchmarks.
- PLM(G): Our proposed method from Equation 4 employed with the Gaussian kernel. The normalization parameter k (the number of dividing intervals of the paragraph where the kernel values are computed) was set to 20. (We experimented with a few other settings for k which resulted in similar performance.) Passage scores were further smoothed with the document score, as in Equation 2, with $\lambda = 0.9$. We report the performance for several steepness (σ) values.
- PLM(Tra): Using PLM with the trapezoid kernel, $k = 20$, and document score smoothing ($\lambda = 0.9$). Performance is reported for several steepness (σ) values.

4.2 Results

Table 1 shows the performance numbers of the different passage scoring methods for the focused task benchmarks of 2009 and 2010. The numbers on the first row are for the

[1] While the focused task encourages systems to retrieve short passages, we decided to experiment with Wikipedia paragraphs as our basic retrievable units. The retrieval of other passage types, e.g., sentences or window-based, is left for future research.

Method	2009 MAiP	2009 iP[.01]	2009 iP[.1]	2010 MAiP	2010 iP[.01]	2010 iP[.1]
INEX best run	0.19	0.63	0.41			
psg	0.09	0.54	0.26	0.06	0.34	0.17
psgDoc	0.22	0.61	0.46	0.15	0.50	0.37
psgNeighbor	0.22	0.63	0.46	0.15	0.50	0.37
PLM(G)						
$\sigma = 500$	0.23	$\underline{0.67}$	$\underline{0.51}$	0.16	$\underline{0.56}$	$\underline{0.41}$
$\sigma = 1000$	0.24	$\underline{0.68}$	$\underline{0.53}$	0.17	$\underline{0.57}$	$\underline{0.43}$
$\sigma = 2000$	$\underline{\mathbf{0.25}}$	$\underline{\mathbf{0.72}}$	$\underline{\mathbf{0.54}}$	$\underline{\mathbf{0.18}}$	$\underline{\mathbf{0.57}}$	$\underline{\mathbf{0.44}}$
$\sigma = 3000$	$\underline{0.25}$	$\underline{0.72}$	$\underline{0.54}$	$\underline{0.18}$	$\underline{0.57}$	$\underline{0.44}$
$\sigma = 10000$	0.24	$\underline{0.70}$	$\underline{0.53}$	0.17	$\underline{0.54}$	$\underline{0.42}$
PLM(Tra)						
$\sigma = 1000$	0.23	$\underline{0.67}$	0.49	0.16	$\underline{0.53}$	0.39
$\sigma = 10000$	0.24	$\underline{0.71}$	$\underline{0.52}$	0.17	$\underline{0.56}$	$\underline{0.44}$
$\sigma = 1e5$	$\underline{\mathbf{0.25}}$	$\underline{\mathbf{0.72}}$	$\underline{\mathbf{0.53}}$	$\underline{\mathbf{0.18}}$	$\underline{\mathbf{0.56}}$	$\underline{\mathbf{0.44}}$
$\sigma = 1e6$	$\underline{0.25}$	$\underline{0.72}$	$\underline{0.53}$	0.17	$\underline{0.56}$	$\underline{0.44}$
$\sigma = 1e7$	0.23	0.63	$\underline{0.51}$	0.16	0.51	0.40

Table 1: The performance of the passage scoring methods over the INEX focused task benchmarks of 2009 and 2010. Boldface marks the best result in a column. Statistically significant improved results over psgDoc are underlined.

best INEX run for 2009 as measured by the *iP[.01]* official measure[2].

We can see in Table 1 that the performance of psg is very low compared to that of psgDoc. This finding is in accordance with previous studies that showed that passage relevance cannot be effectively estimated without considering the passage context [7]. Moreover, psgDoc performs equivalently to the best run in INEX 2009, which is a strong baseline to compare with. We can also see that the performance of psgNeighbor is almost identical to that of psgDoc. Generalizing this approach by considering $k > 1$ passages from left and right did not gain any improvement. This finding implies that using the neighbor passages of the passage at hand for score smoothing has no additional benefit on top of using the document score for smoothing.

Both PLM methods yield substantial, and (almost always) statistically significant, improvement over psgDoc when setting σ to large values. The largest gap in performance is observed in *iP[.01]* which measures the precision of topmost results, at 0.01 recall level. The performance differences between the two kernels were found to be statistically insignificant. Interestingly, the optimal performance for PLM was achieved when setting the free contextualization parameter (σ) to a large value for both kernels ($\sigma = 2000$ for PLM(G) and $\sigma = 1e5$ for PLM(Tra)). Note that the larger σ is, the higher the contextualization is, i.e., all query term offsets, regardless of where they appear in the document, contribute to the passage score. On the other hand, when setting the steepness parameter to an extremely large value, performance drops for both kernels. The conclusion is that contextualization based on term positions is important for passage retrieval. That is, query-terms occurrence in a document should affect the relevance estimation of passages, even if the passages are quite far from the occurrence positions. Yet, this effect should (marginally) decay with the distance.

[2] In 2010 the focused task was modified; participants were asked to retrieve only 1000 characters per topic, hence the precision results of INEX participants are extremely low, and cannot be compared to those of our method which is based on retrieving full paragraphs.

5. SUMMARY

We presented a novel contextualization approach for passage retrieval. All query term occurrences in a document affect the score assigned to a passage. The effect depends on the proximity of the occurrences of the query terms to the passage. Empirical evaluation showed that our approach posts performance that is substantially better than that of previously proposed passage-based retrieval methods that use various contextualization approaches.

6. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. UMASS at TREC 2004 — novelty and hard. In *Proceedings of the Thirteenth Text Retrieval Conference (TREC-13)*, 2004.

[2] J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of SIGIR*, pages 314–321, 2003.

[3] P. Arvola, S. Geva, J. Kamps, R. Schenkel, A. Trotman, and J. Vainio. Overview of the INEX 2010 ad hoc track. *Comparative Evaluation of Focused Retrieval*, pages 1–32, 2011.

[4] P. Arvola, M. Junkkari, and J. Kekäläinen. Generalized contextualization method for XML information retrieval. In *Proceedings of CIKM*, pages 20–27, 2005.

[5] M. Beigbeder. Focused retrieval with proximity scoring. In *Proceedings of SAC*, pages 1755–1759, 2010.

[6] R. Blanco and H. Zaragoza. Finding support sentences for entities. In *Proceedings of SIGIR*, pages 339–346, 2010.

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

[8] O. de Kretser and A. Moffat. Effective document presentation with a locality-based similarity heuristic. In *Proceedings of SIGIR*, pages 113–120, 1999.

[9] R. T. Fernández, D. E. Losada, and L. A. Azzopardi. Extending the language modeling framework for sentence retrieval to include local context. *Inf. Retr.*, 14(4):355–389, Aug. 2011.

[10] S. Geva, J. Kamps, M. Lethonen, R. Schenkel, J. Thom, and A. Trotman. Overview of the INEX 2009 ad hoc track. *Focused Retrieval and Evaluation*, pages 4–25, 2010.

[11] M. Kaszkiel and J. Zobel. Passage retrieval revisited. In *Proceedings of SIGIR*, pages 178–185, 1997.

[12] Y. Lv and C. Zhai. Positional language models for information retrieval. In *Proceedings of SIGIR*, pages 299–306, 2009.

[13] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL*, page 20, 2004.

[14] V. Murdock and W. B. Croft. A translation model for sentence retrieval. In *Proceedings of HLT/EMNLP*, 2005.

[15] M. A. Norozi, P. Arvola, and A. P. de Vries. Contextualization using hyperlinks and internal hierarchical structure of wikipedia documents. In *Proceedings of CIKM*, pages 734–743, 2012.

[16] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Proceedings of TREC-3*, 1994.

[17] G. Salton, J. Allan, and C. Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of SIGIR*, pages 49–58, 1993.

[18] S. Tellex, B. Katz, J. Lin, A. Fernandes, and G. Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of SIGIR*, pages 41–47. ACM, 2003.