OUC's participation in the 2010 INEX Book Track

Michael Preminger[1] and Ragnar Nordlie[1]

Oslo University College

Abstract. In this article we describe the Oslo University College's participation in the INEX 2010 Book track. The OUC has submitted retrieval results for the "prove it" task with traditional relevance detection combined with some rudimental detection of confirmation. We call for a broader discussion of a more meaning-oriented (semantics-aware) approach to retrieval in digitized books, with the "prove it" task (classifiable as a simple semantics- aware retrieval activity) providing the INEX milieu with a suitable context to start this discussion.

1 Introduction

In recent years large organizations like national libraries, as well as multinational organizations like Microsoft and Google have been investing labor, time and money in digitizing books. Beyond the preservation aspects of such digitization endeavors, they call on finding ways to exploit the newly available materials, and an important aspect of exploitation is book and passage retrieval.

The INEX Book Track[1], which has been running since 2007, is an effort aiming to develop methods for retrieval in digitized books. One important aspect here is to test the limits of traditional methods of retrieval, designed for retrieval within "documents" (such as news-wire), when applied to digitized books. One wishes to compare these methods to book-specific retrieval methods.

One important mission of such retrieval is supporting the generation of new knowledge based on existing knowledge. The generation of new knowledge is closely related to access to – as well as faith in – existing knowledge. One important component of the latter is claims about facts. This year's "prove it" task, may be seen as challenging the most fundamental aspect of generating new knowledge, namely the establishment (or refutal) of factual claims encountered during research.

On the surface, this may be seen as simple retrieval, but proving a fact is more than finding relevant documents. This type of retrieval requires from a passage to "make a statement about" rather than "be relevant to" a claim, which traditional retrieval is about. The questions we pose here are:

- *what is the difference between simply being relevant to a claim and expressing support for a claim*
- how do we modify traditional retrieval to reveal support or refutal of a claim?

We see proving and denial of a statement as different tasks, both classifiable as semantics-aware retrieval, suspecting that the latter is a more complicated task. This paper attempts at applying some rudimentary techniques of detecting the confirmation (proving) of a statement. The rest of the paper discusses these tasks in the context of meaning-oriented retrieval in books.

2 Indexing and retrieval strategies

The point of departure of the strategies discussed here is that confirming or refuting a statement is a simple action of speech that does not require from the book (the context of the retrieved page) to be ABOUT the topic covering the fact. This means that we do not need the index to be context-faithful (pages need not be indexed in a relevant book context). It is more the formulation of the statement in the book or page that matters. This is why we need to look for words (or sequences of words) or sentences that indicate the stating of a fact. A simple strategy is looking for the occurrence of words like "is", "are", "have", "has" a.s.o, that, in combination with nouns from the query (or fact formulation), indicate a possible act of confirming the fact in question.

Further focus may be achieved by detecting sentences that include (WH-) question indicators or a question-mark and pruning these from the index, so that pages that only match the query through such sentences are omitted or weighed down during retrieval.

Against this background we were trying to construct runs that emphasized pages that are confirmative in style. We attempted to divide the pages in the collection into categories of how confirmative they are, and indexed them individually (each page comprising a document). Occurrences of the words *is, are, was, were, have, has* were counted in each page, and a ratio between this sum and the total number of words in the page was calculated. Based on a sample of the pages, three levels were defined, so that pages belonging to each of the levels were assigned a tag, accordingly. It may be argued that a richer set of confirmation indicators could be applied. Our claim is that the selected words should function as style indicators, not as content indicators (the content catered for by the topic, i.e. the factual claim under scrutiny), and were therefore sufficient. A larger collection could incur noise.

These tags then facilitated weighting pages differently (based on their proportion of confirmatory words) when retrieving candidates of confirming pages. Retrieval was performed using the Indri language model in the default mode, weighting the confirmatory pages differently, as indicated above. As the primary aim here is to try and compare traditional retrieval with "prove it", there was no particular reason to divert from the default.

The 2010 tasks have been featuring a relatively large number of topics (83). As a new method was employed for collecting relevance assessments, only 21 of these queries were ready for evaluation at deadline time, and it is these queries that are used for retrieval.

3 Runs and Results

Based on the index that we were constructing, we weighted relevant pages on two levels: pages that featured 1 percent or more confirmatory words, and pages that featured 3 percents or more confirmatory words (the latter including the former) were weighted double, quintuple (5x) and decuple (10x) the baseline. Our baseline was normal, non-weighted retrieval, as we would do for finding relevant pages. We were using the indri combine operation with no changes to the default setting (regarding smoothing, a.s.o).

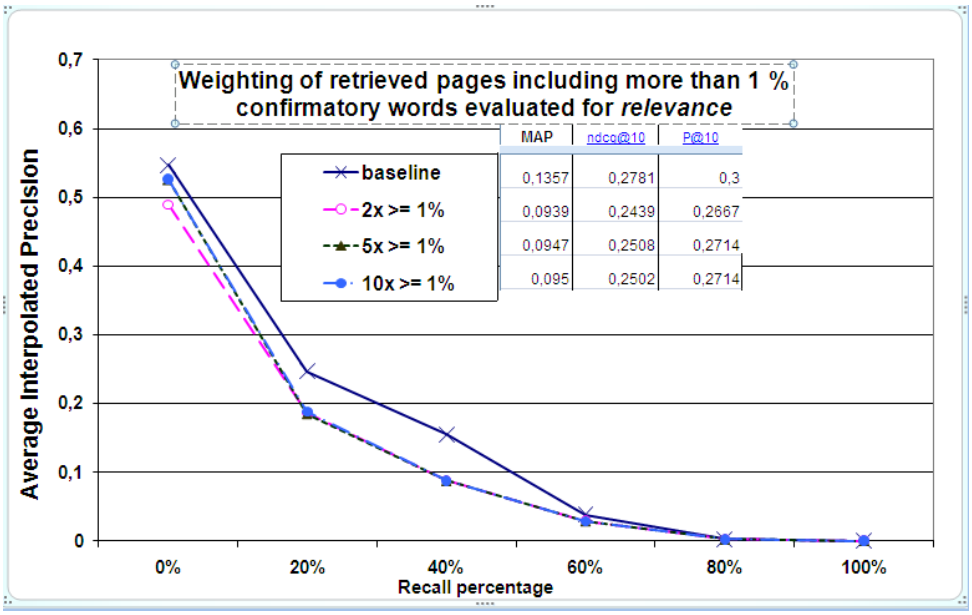
The analysis was carried out twice: once against the entire official qrel file (meaning that all assessed book pages judged either confirming/refuting – or merely relevant to – the statement-query are taken to be relevant (Figure 1). The second analysis was done against a filtered version of the official qrel file, featuring only the pages assessed as confirming/refuting (Figure 2).

The purpose was to see if the rate of confirmatory words can be used as a "prove it" indicator, given that the relevance assessments properly reveal pages that confirm the factual statement. Weighting retrieved pages that feature 1 percent or more confirmatory words does not seem to outperform the baseline at any weighting level, at any region of the precision recall curve (Subfigures 1(a) and 2(a)). The reason for that may be that quite many pages belong to this category. The weighting thus seems to hit somewhat randomly. An occurrence rate of confirmatory words of one percent seems not to discriminate "proving" pages.

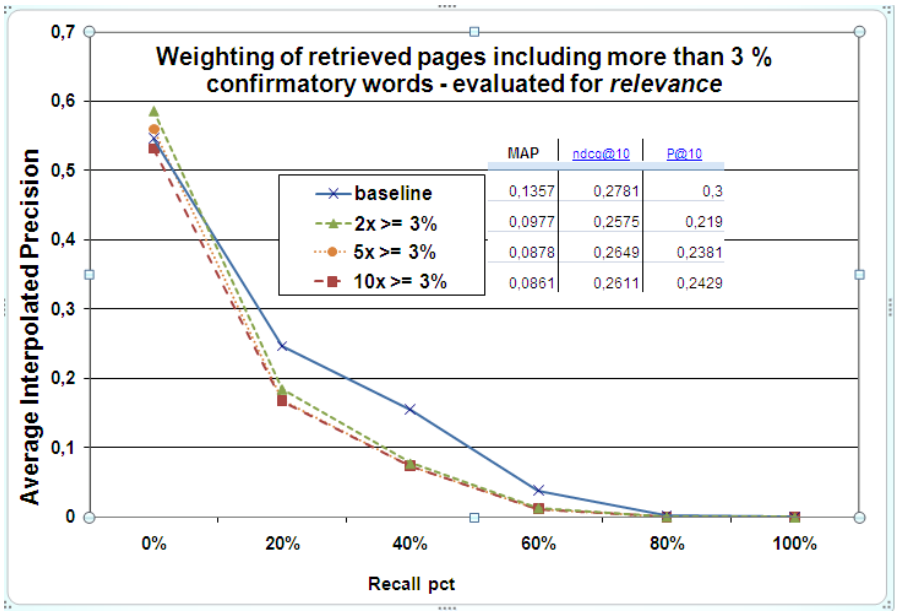
The results are more interesting when restricting the weighting to pages that feature 3 percents or more confirmatory words. Here the results are different at the low recall and the high recall regions. Directing our attention to the low recall region first, we see in Figure 1(b) that both doubling the weight and, to a lesser extent quintupling it, slightly outperform the the baseline. The effect is a bit clearer when evaluating by pages assessed as confirming (Figure 2(b)). Here also decupling the weight given to pages with 3 percent or more of confirmatory words slightly outperforms the baseline.

No treatment seems to outperform the baseline in the higher recall regions. Subject, of course, to a more thorough scrutiny of the result, this could indicate that collecting many books that prove a statement is not likely to be better supported by this approach than by traditional retrieval, whereas only finding very few such books (early hits) might benefit from it. The reason for that may be approached by looking at single relevant pages retrieved at the low and high recall regions. This kind of treatment was beyond the scope of the present paper. The value of pursuing it may be limited in light of the overall results.

Looking at the "prove it" task in terms of traditional retrieval, the temporary conclusion would be that the treatment experimented with here may be in the right direction, and further pursuit of it has some potential of good retrieval, particularly if it is the low recall region that is important (early hits). If the

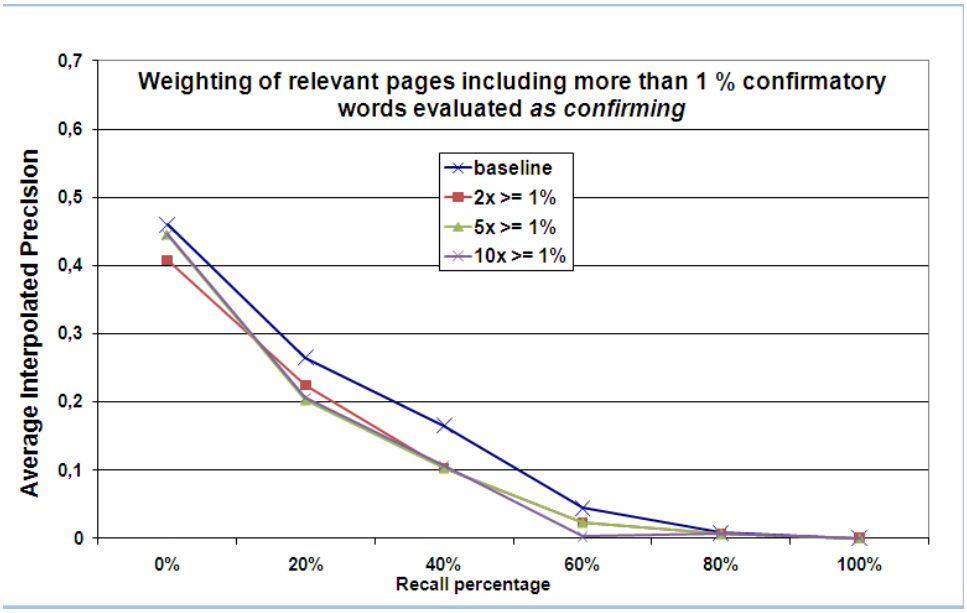


(a) Weighting relevant pages with 1 percent or more confirmatory words

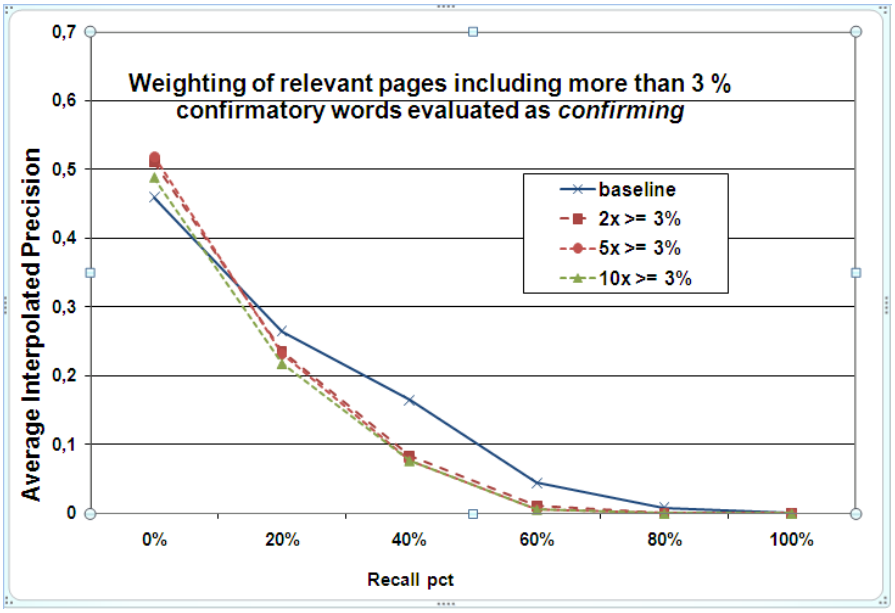


(b) Weighting relevant pages with 3 percent or more confirmatory words

Fig. 1. Precision-recall curves for detecting relevant pages. Baseline marked by solid lines in both subfigures.



(a) Weighting relevant pages with 1 percent or more confirmatory words



(b) Weighting relevant pages with 3 percent or more confirmatory words

Fig. 2. Precision-recall curves for detecting confirming (proving) pages. Baseline marked by solid lines in both subfigures.

purpose is collecting as many books as possible as evidence for a claim then the approach does not seem as promising.

4 Discussion

Utilizing digital books poses new challenges on information retrieval. The mere size of the book text poses both storage, performance and content related challenges as compared to texts of more moderate size. But the challenges are even greater if books are to be exploited not only for finding facts, but also to support exploitation of knowledge, identifying and analyzing ideas, a.s.o.

For example, we suspect that confirming and refuting a factual statement, the Book Track 2010 "prove it" task, both belong to a class of activities that extend the current scope of information retrieval. The notion of relevance is a well known challenge in IR [2]. We suspect that the "prove it" notion is by no means simpler. Confirming a fact may have many facets, based on how complicated the fact is. A fact like: *The ten tribes forming the northern kingdom of Israel (aka the ten lost tribes) disappeared after being driven to exile by the Assyrians, several hundreds years before Christ* (topic 2010003) may be confirmed on several levels. Should all minor details be in place for the fact to be confirmed? What if the book states that it was the Babylonians, rather than the Assyrians who sent the tribes into exile, the rest of the details being in agreement with the statement: is the fact then confirmed? Moreover, detecting the refutal of a statement is arguably a totally different activity than detecting its confirmation. This poses challenges not only to mere retrieval, but also to its evaluation, at all levels.

Even though such activities may be developed and refined using techniques from e.g. Question Answering[3], we suspect that employing semantics-aware retrieval [4,5], which is closely connected to the development of the Semantic Web [6] would be a more viable (and powerful) path to follow.

Within the INEX Book track, the "prove it" task can thus serve as a splendid start of a broader discussion around detecting meaning rather than only matching strings. Many projects under way are already using ontologies to aid in tagging texts of certain kinds (e.g. philosophical essays)[7] to indicate certain meaning, with the aim of supporting the analysis of these texts. Is this a viable task for the INEX Book track? Is it a viable path for information retrieval?

5 Conclusion

This article is an attempt to start a discussion about semantics-aware retrieval in the context of the INEX book track. Proving of factual statements is discussed in light of some rudimental retrieval experiments incorporating the detection of confirmation (proving) of statement. We also discuss the task of proving statement, raising the question whether it is classifiable as a semantics-aware retrieval task.

References

1. Kazai, G., Koolen, M., Landoni, M.: Summary of the book track. In: INEX 2009
2. Mizzaro, S.: Relevance: The whole history. Journal of the American Society of Information Science **48**(9) (1997) 810–832
3. VOORHEES, E.M.: The trec question answering track. Natural Language Engineering **7** (2001) 361–378
4. Tim Finin, James Mayfield, A.J.R.S.C., Fink, C.: Information retrieval and the semantic web. In: Proc. 38th Int. Conf. on System Sciences, Digital Documents Track (The Semantic Web: The Goal of Web Intelligence)
5. Mayfield, J., Finin, T.: Information retrieval on the semantic web: Integrating inference and retrieval. In: SIGIR Workshop on the Semantic Web, Toronto
6. Berners-Lee, T., H.J., Lassila, O.: The semantic web. Scientific American (2001)
7. Zllner-Weber, A.: Ontologies and logic reasoning as tools in humanities? DHQ: Digital Humanities Quarterly **3**(4) (2009)