OUC's participation in the 2011 INEX Book Track

Michael Preminger[1] and Ragnar Nordlie[1]

Oslo and Akershus University College of Applied Science

Abstract. In this article we describe the Oslo University College's participation in the INEX 2011 Book track. In 2010, the OUC submitted retrieval results for the "Prove It" task with traditional relevance detection combined with some rudimental detection of confirmation. In line with our belief that proving or refuting facts are different semantic aware actions of speech, we have this year attempted to incorporate some rudimentary semantic support based on the WordNet database.

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 posed in 2010 were:

- 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 also made the claim that ”Prove It” sorts within the (not very well-defined) category ”semantic-aware retrieval”, which, for the time being will be defined by us as retrieval that goes beyond simple string matching, and is aware of the meaning (semantics) of text.

Those question, being rhetorical in part, may be augmented by the questions

- How can one detect the meaning of texts (words, sentences and passages) and incorporate those in the retrieval process to attain semantic-aware retrieval

and consequently

- can one exploit technologies developed within the semantic web to improve semantic-aware retrieval

The latter is not directly addressed in this paper, but we claim that the techniques used here point in this direction.

2 The ”Prove It” Task

2.1 Task Definition and User Scenario

The prove-it task is still at its infancy, and may be subject to some modifications in the future. Quoting the user scenario as formulated by the organizers

> The scenario underlying this task is that of a user searching for specific information in a library of books that can provide evidence to confirm or refute a given factual statement. Users expect to be pointed directly at book pages that can help them to confirm or refute the claim of the topic. Users are assumed to view the ranked list of retrieved book pages starting from the top of the list and moving down, examining each result. No browsing is considered (only the returned book pages are viewed by users).

This user scenario is a natural point of departure as it is in the tradition of information retrieval and facilitates the development of the task by using existing knowledge. As a future strategy, it may be argued that this user scenario is gradually modified, as ranking in the context of proving is a highly complex process, and, in the context where Prove-it algorithms are most likely to be used, arguably superfluous.

2.2 What Is a Proof?

What constitutes a proof is well defined in fields like mathematics and computer science. In connection with a claim or a statement of fact, it is less obvious what demands a passage of text should satisfy in order to be considered proof of the claim. Obviously, we are looking for a passage which expresses a relevant truth about the claim, but what are the characteristics which signal a sufficient degree of relevance and truthfulness? We might want to identify a trustworthy passage, which in its turn might be identified by considering the source of the passage, the degree to which the passage agreed with other passages treating the same claim or fact, or the centrality of the claim to the main content of the text. We might want to identify a concentrated passage, a passage where the largest amount of elements contained in the claim were represented or where they were by some measure most heavily represented. We might look for a definitional passage, which typographically or linguistically showed the characteristics of a definition. Or we might try to identify a ”proof” by linguistic characteristics, mostly semantic, which might be of different kinds: certain typical words might be relatively consistently used to speak about a fact or claim in a ”proving” manner, writing in a ”proving” mode might entail using terms on a certain level of specificity, etc. These latter aspects are orthogonal to the statement or claim itself in the sense that they (at least ideally) apply equally to whatever claim being the subject of proving / confirming.

2.3 Semantic Approaches to Proof

A statement considered as a ”proof” (or confirmation) may be characterized semantically by several indicators:

- the phenomenon to be supported may be introduced or denoted by specific terms, for instance verbs indicating a definition: ”is”, ”constitutes”, ”comprises” etc.
- terms describing the phenomenon may belong to a specific semantic category
- nouns describing the phenomenon may be on a certain level of specificity
- verbs describing the phenomenon may denote a certain type of action or state

Deciding which specificity level or which semantic categories will depend on the semantic content and the relationship between the terms of the original claim. Without recourse to the necessary semantic analysis, we assume that in general, terms indicating a proof / confirmation will be on a relatively high level of specificity. It will in some way constitute a treatment of one or more aspects of the claim at a certain level of detail, which we expect to be reflected in the terminology which is applied.

As an initial exploration of these potential indicators of proof, without access to semantic analysis of the claim statements, we are investigating whether terms, in our case nouns, found on a page indicated as a potential source of proof diverges in a significant way from other text in terms of level of specificity. We determine the level of noun specificity through their place in the WordNet([2]) term hierarchies.

As stated further down the paper, this is an initial use of this type of semantics in retrieval, and the only thing we can hope for is that it gives us an indication about whether proceeding in this path is viable.

2.4 Ranking According to "Proof Efficiency"?

In this paper we are still following the two-step strategy of first finding pages relevant to the claim, and from those pages trying to identify pages that are likely to prove the claim[1]. The first step is naturally done using current strategies for ranked retrieval. The second stage identifies *among relevant documents* those which prove / confirm the statement. Rank order is not necessarily preserved in this process: if document A comprises a better string-wise match with the claim than does document B, document B can still be more efficient at proving the claim than document A is. Not all elements that make a document relevant also make it a good prover

Another issue is the context in which prove-it is used. One example is the writing of a paper. A writer is (again, arguably) more likely to evaluate a greater number of sources for proof of a claim than he or she would in a context of pure fact finding. Additionally, different contexts would arguably invite different proof emphases. All this advocates for use of other strategies of presenting proving results than ranked lists.

3 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. In this way the "Prove It" task is different than e.g. the one referred to in [3] This means that we do not need the index we build for search purposes to be context-faithful (pages need not be indexed in a relevant book context). It is the formulation of the statement in the book or page that matters.

3.1 Indexing

In line with the above, indexing should facilitate two main aspects at retrieval time: identifying relevant pages and finding which of these is likely to prove a claim. The first aspect is catered for creating a simple index of all the words in the corpus, page by page. The pages are treated as separate documents regardless of the book in which they appear. The second aspect is catered for by calculating the average specificity of each page and tagging each page by one of a number of specificity tags. The latter are determined as described in Section 3.2

3.2 Calculating Specificity

At this stage of the research, the aspect of finding pages likely to prove a claim is catered for by statistically measuring the average specificity of words that occur in the page. We do this by calculating the specificity of each word and then averaging the measure of specificity of all the words in a page, as described

[1] We see refutal as a totally different type of task and will not address it in this paper.

below. To accomplish that, we have augmented the WordNet database (ref) by a Direct Acyclic Graph (DAG) of all the nouns, which lets us calculate a relative specificity of each word by its average position in this graph. Words closer to the common root of the graph (measured as a number of steps) are less specific, whereas words closer to the leaves are more specific. For each word in a trajectory, the specificity S is calculated as

$$S = \frac{P}{L},$$

where P is the position of the words in the trajectory (number of steps away from the root) and L is the length of the trajectory from root to leaf. Since this is a graph and not a tree, each word (a string of characters), even a leaf, may belong to more than one trajectory depending on the number of senses / synsets it participates in, and the number of parallel synsets it is a descendent of. Since we generally cannot know which sense of a word a certain occurrence stands for, we assign to each word (string of characters) the average of its specificities. Each page is then assigned the average of the specificities of its constituent words. Words not in the graph are assigned the "neutral" value of 0.5.

The pages are then categorized into predefined intervals of average specificity. We were working with an interval resolution of 5%, where pages between x and $x + 5\%$ are categorized together for each $x = 5\%, 10\%, 15\%$.... Each interval has its own tag for indexing purposes. These tags then facilitate weighting pages differently at retrieval time when retrieving candidates of confirming pages.

4 Runs and Results

We look at results in two different sub-scenarios. Instant - to what extent the system supports "instant proving" of documents. In this sub-scenario the first document that proves the statement is taken as the statement's proof, and no further pages are visited. This mode is well represented by the MRR (Mean reciprocal rank) measure. Thorough - more pages are visited to establish the proof of the statement. This is well represented by the MAP measure, and precision-recall curves. The NDCG (official measure of the Track) expresses both sub scenarios.

The way we measure the effect of specificity is that we, at retrieval time, boost up pages with different rates of specificity (as measured and tagged in 3.2) weighting them up by different factors. We operate with two range-modes:

- A narrow specificity interval (5 percent points between $x\%$ and $x + 5\%$) (*eq*) "spec_2x_eq_55" means that pages with a specificity between 55 and 60 are weighted twice as much as other pages.
- A one-sided specificity interval greater than or equal to an interval point (*ge*). "spec_5x_ge_55" means that the pages with a specificity equal to or greater than 55 are given five time the weight of lower-specificity pages at retrieval time.

In this section we present two types of runs:

- Calibration runs, runs that are meant to find good parameter candidates for specificity and document weighting at retrieval time.
- Full scale performance runs

4.1 Calibration Runs

The calibration runs are runs performed against an index of this subset of the pages only containing the pages appearing in the recall base of at least one of the topics (the pages in the applicable *.qrel file). The performance runs are runs against the entire page corpus.

The purpose of the calibration runs is to more sensitively (and more effortlessly) measure the effect of the parameters and combination of them on several performance indicators, before applying the best performing parameters to the full-scale performance runs. An index is constructed, containing only the pages appearing in the ”qrel” files, giving the algorithms fewer non-relevant pages to deal with. A number of pre-runs not reported here have indicated an effective range of specificity (just above the neutral 0.5 rate) that perform better than both lower and higher measures.

In figure 1 we can see that the intermediate two-sided ranges, 55% and 60%, generally perform better than the 50% 65% ranges. Narrow (two-sided) ranges perform better than one-sided. Based on these results, a specificity rate of 60% gives the best reciprocal mean rank measure, meaning that the document performing best is in average second or third in the ranking list. A slightly lower specificity rate (55%) seems to better support the sub scenario where the user looks at a number of pages before accepting a statement as confirmed (as expressed by the map and ndcg measures).

4.2 Performance Runs

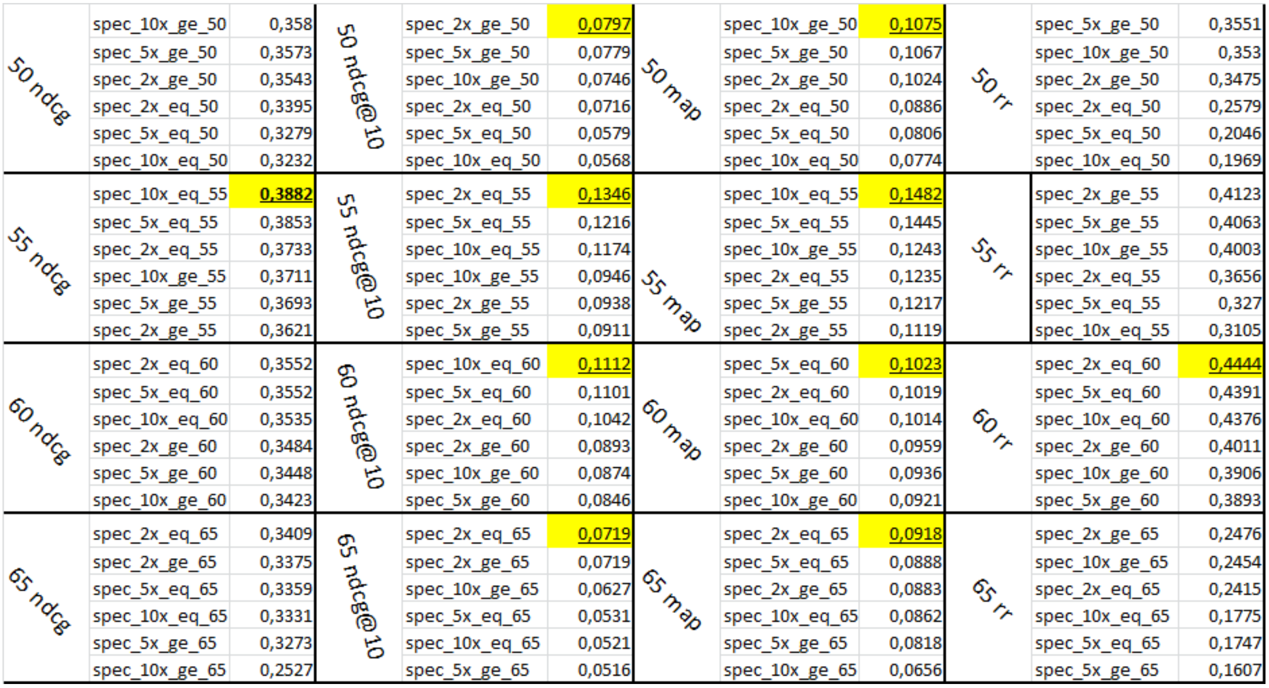
In figure 2 we present full-scale runs made against the full-scale index (17M pages), using the best parameters of the calibration runs.

The results presented here are an attempt at relating this year’s results to our 2010 results [4]. Figure 3 shows the results of weighting pages featuring 3 percents or more confirmatory words at retrieval time, weighted double, quintuple (5x) and decuple (10x) the baseline[2]. We do spot a slight improvement in the 2011 results, but it is hard to say whether it is significant.

5 Discussion, Limitation and Further Research

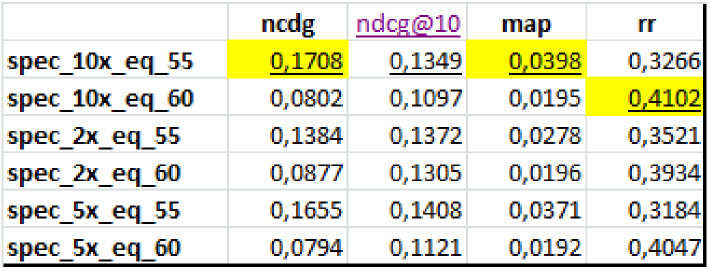
At the same time that the book world becomes more and more digital,as old books are being digitized and new books are increasingly published digitally,

[2] For these, as well as all other plots, We were using the indri combine / weight operation (a combination of weighted expression) with no changes to the default setting (regarding smoothing, a.s.o),



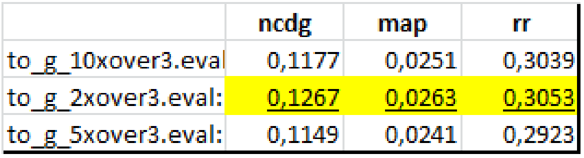
50 ndcg	spec_10x_ge_50	0,358	50 ndcg@10	spec_2x_ge_50	0,0797	50 map	spec_10x_ge_50	0,1075	50 rr	spec_5x_ge_50	0,3551
	spec_5x_ge_50	0,3573		spec_5x_ge_50	0,0779		spec_5x_ge_50	0,1067		spec_10x_ge_50	0,353
	spec_2x_ge_50	0,3543		spec_10x_ge_50	0,0746		spec_2x_ge_50	0,1024		spec_2x_ge_50	0,3475
	spec_2x_eq_50	0,3395		spec_2x_eq_50	0,0716		spec_2x_eq_50	0,0886		spec_2x_eq_50	0,2579
	spec_5x_eq_50	0,3279		spec_5x_eq_50	0,0579		spec_5x_eq_50	0,0806		spec_5x_eq_50	0,2046
	spec_10x_eq_50	0,3232		spec_10x_eq_50	0,0568		spec_10x_eq_50	0,0774		spec_10x_eq_50	0,1969
55 ndcg	spec_10x_eq_55	**0,3882**	55 ndcg@10	spec_2x_eq_55	0,1346	55 map	spec_10x_eq_55	0,1482	55 rr	spec_2x_ge_55	0,4123
	spec_5x_eq_55	0,3853		spec_5x_eq_55	0,1216		spec_5x_eq_55	0,1445		spec_5x_ge_55	0,4063
	spec_2x_eq_55	0,3733		spec_10x_eq_55	0,1174		spec_10x_ge_55	0,1243		spec_10x_ge_55	0,4003
	spec_10x_ge_55	0,3711		spec_10x_ge_55	0,0946		spec_2x_eq_55	0,1235		spec_2x_eq_55	0,3656
	spec_5x_ge_55	0,3693		spec_2x_ge_55	0,0938		spec_5x_ge_55	0,1217		spec_5x_eq_55	0,327
	spec_2x_ge_55	0,3621		spec_5x_ge_55	0,0911		spec_2x_ge_55	0,1119		spec_10x_eq_55	0,3105
60 ndcg	spec_2x_eq_60	0,3552	60 ndcg@10	spec_10x_eq_60	0,1112	60 map	spec_5x_eq_60	0,1023	60 rr	spec_2x_eq_60	0,4444
	spec_5x_eq_60	0,3552		spec_5x_eq_60	0,1101		spec_2x_eq_60	0,1019		spec_5x_eq_60	0,4391
	spec_10x_eq_60	0,3535		spec_2x_eq_60	0,1042		spec_10x_eq_60	0,1014		spec_10x_eq_60	0,4376
	spec_2x_ge_60	0,3484		spec_2x_ge_60	0,0893		spec_2x_ge_60	0,0959		spec_2x_ge_60	0,4011
	spec_5x_ge_60	0,3448		spec_10x_ge_60	0,0874		spec_5x_ge_60	0,0936		spec_10x_ge_60	0,3906
	spec_10x_ge_60	0,3423		spec_5x_ge_60	0,0846		spec_10x_ge_60	0,0921		spec_5x_ge_60	0,3893
65 ndcg	spec_2x_eq_65	0,3409	65 ndcg@10	spec_2x_eq_65	0,0719	65 map	spec_2x_eq_65	0,0918	65 rr	spec_2x_ge_65	0,2476
	spec_2x_ge_65	0,3375		spec_2x_ge_65	0,0719		spec_5x_eq_65	0,0888		spec_10x_ge_65	0,2454
	spec_5x_eq_65	0,3359		spec_10x_ge_65	0,0627		spec_2x_ge_65	0,0883		spec_2x_eq_65	0,2415
	spec_10x_eq_65	0,3331		spec_5x_eq_65	0,0531		spec_10x_eq_65	0,0862		spec_10x_eq_65	0,1775
	spec_5x_ge_65	0,3273		spec_10x_eq_65	0,0521		spec_5x_ge_65	0,0818		spec_5x_eq_65	0,1747
	spec_10x_ge_65	0,2527		spec_5x_ge_65	0,0516		spec_10x_ge_65	0,0656		spec_5x_ge_65	0,1607

Fig. 1. Calibration runs: NDCG, MAP and Mean reciprocal rank results for runs using different parameter values



	ncdg	ndcg@10	map	rr
spec_10x_eq_55	0,1708	0,1349	0,0398	0,3266
spec_10x_eq_60	0,0802	0,1097	0,0195	0,4102
spec_2x_eq_55	0,1384	0,1372	0,0278	0,3521
spec_2x_eq_60	0,0877	0,1305	0,0196	0,3934
spec_5x_eq_55	0,1655	0,1408	0,0371	0,3184
spec_5x_eq_60	0,0794	0,1121	0,0192	0,4047

Fig. 2. Performance runs: NDCG, MAP and Mean reciprocal rank results for runs against a full scale index



	ncdg	map	rr
to_g_10xover3.eval	0,1177	0,0251	0,3039
to_g_2xover3.eval:	0,1267	0,0263	0,3053
to_g_5xover3.eval:	0,1149	0,0241	0,2923

Fig. 3. The 2010 results: NDCG, MAP and Mean reciprocal rank results for runs against a full scale index

information not published in book format becomes more and more "semantic" in the sense that data pieces (as opposed to exclusively documents in the web's first years) are linked together and made available. These two parallel development entail great opportunities in the exploitation of book material for different purposes, of which the topic of this paper is one example.

This paper provides an example of the possibilities and the challenges. Whereas "WordNet specificity", here representing content independent linguistic semantic, is one simple example of information that can be used to systematically extract semantics from written content, other much larger and much more complicated sources of semantics, the semantic web and linked data, are waiting to be used in a similar (or related) way. To explore these possibilities we will need to experiment with more modern texts than what our present test collection contains.

To judge by the results of the runs presented here, this path of research, though promising, still requires a lot of modification and calibration.

Exploring the semantics of a page in a basically statistical manner may be seen as a superposition of independent components. Counting occurrences of special words is one component on which we superimpose the detection of noun specificity. The treatment using WordNet represents further progress from the 2010 experiments, but is still rudimentary. Nouns are currently the only word-class we are treating, using only level of specificity. trying to detect classes nouns using the lateral structure of synsets may be another path to follow. It is also conceivable that treating of other word classes, primarily verbs, might contribute to the treatment. Verbs are more complicated than nouns in WordNet and such treatment will be more demanding.

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.

This article represents work in progress. We explore techniques gradually in an increasing degree of complexity, trying to adapt and calibrate them.

Even though such activities may be developed and refined using techniques from e.g. Question Answering[5], we suspect that employing semantics-aware retrieval [6,7], which is closely connected to the development of the Semantic Web [8] would be a more viable (and powerful) path to follow.

One obstacle particular to this research is the test collection. Modern ontologies code facts that are closely connected to the modern world. For example the Yago2 [9] ontology, that codes general facts automatically extracted from Wikipedia, may be complicated to apply to an out-of-copyright book collection emerging from academic specialized environments. But this is certainly a path to follow.

6 Conclusion

This article is a further step in a discussion about semantics-aware retrieval in the context of the INEX book track. Proving (or confirmation or support) of factual statements is discussed in light of some rudimental retrieval experiments incorporating semantics. We also discuss the task of proving statement, raising the question whether it is classifiable as a semantics-aware retrieval task. Results are highly inconclusive.

References

1. Kazai, G., Koolen, M., Kamps, J., Doucet, A., Landoni, M.: Overview of the inex 2010 book track: Scaling up the evaluation using crowdsourcing. In: Comparative Evaluation of Focused Retrieval. Volume 6932 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2011) 98–117
2. Fellbaum, C.: WordNet : an electronic lexical database. MIT Press, Cambridge, Mass (1998)
3. Cartright, M.A., Feild, H., Allan, J.: Evidence finding using a collection of books. In: BooksOnline '11 Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing, Amherst, MA (2011) 11–18
4. Preminger, M., Nordlie, R.: Ouc's participation in the 2010 inex book track. In: Comparative Evaluation of Focused Retrieval. Lecture Notes in Computer Science, Springer Berlin / Heidelberg (2010) 164–170
5. VOORHEES, E.M.: The trec question answering track. Natural Language Engineering **7** (2001) 361–378
6. Finin, T., Mayfield, J., Joshi, A., Cost, R.S., 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). (2005)
7. Mayfield, J., Finin, T.: Information retrieval on the semantic web: Integrating inference and retrieval. In: SIGIR Workshop on the Semantic Web, Toronto. (2003)
8. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American (2001)
9. Hoffart, J., Suchanek, F., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Special issue of the Artificial Intelligence Journal (2012)