The Responsa Project: Some Promising Future Directions

Moshe Koppel

Dept. of Computer Science
Bar-Ilan University
Ramat-Gan, ISRAEL

Abstract. We present a very brief review of some of the achievements of the Bar-Ilan Responsa Project during the period of Yaacov Choueka's leadership and discuss some of the directions the project might consider in order to meet ongoing challenges.

Keywords: Responsa Project, information retrieval

1 Introduction

One of the crowning achievements of Yaacov Choueka's illustrious career has been his guidance of the Bar-Ilan Responsa project from a fledgling research project to a major enterprise awarded the Israel Prize in 2008. Much of the early work on the Responsa project ultimately proved to be foundational in the now burgeoning area of information retrieval, the science of searching large digitized corpora for information.

In this paper, I will very briefly review some of the project's achievements and will discuss some of the directions the project might consider in order to meet ongoing challenges. (The reader wishing to read an insider's detailed review of the project's achievements and challenges is referred to (Choueka 1990).)

The Responsa project was initiated by Aviezri Fraenkel in 1963, well before massive searchable text corpora became commonplace. In order to appreciate the challenges faced by researchers involved with the Responsa project in those early days, it is instructive to compare the corpus to the most well-known corpus extant at the time, namely, the Brown corpus developed at Brown University (Kucera & Francis 1967).

The Brown corpus consisted of one million words of text assembled for the purpose of studying language use. The selection of texts included in the corpus reflected its intended purpose. They were chosen in an essentially random manner limited primarily by the constraint that the overall corpus be representative of the relative frequency of each of 15 text genres found in the wild in the year 1961. The first 2000 words of each of 500 texts were used. The corpus was eventually tagged for parts of speech and was accompanied by its own search engine.

Both the Responsa corpus and the Brown corpus were seminal efforts to digitize a corpus for the purpose of information retrieval. However, the Responsa corpus differed from the Brown corpus in three main ways:

1. The texts in the Responsa corpus were in Hebrew.
2. The corpus was intended as a source of information and not merely as a source of exemplars of language use.
3. The documents included in the corpus span a period of thousands of years and are of historical significance.

In the following sections, we will consider the particular technological challenges posed by each of these differences and the ways in which these challenges have been, or might one day be, met. After that we will consider a number of open problems concerning the accuracy and provenance of texts included in the corpus.

2 Hebrew

Hebrew is substantively different than English along a variety of linguistic dimensions. Here we focus specifically on those aspects that create the need for special pre-processing in a searchable corpus.

First of all, Hebrew has a far richer morphology than does English. Many English function words are encoded in Hebrew prefixes and a given normal form (root) has many derivative forms, many of which alter the normal form, rather than merely augmenting it. A user searching for some Hebrew word might thus typically be equally interested in a variety of other words sharing the same normal form. Thus, the need for a tool that can identify the normal form of a given word and, conversely, the variety of derivatives corresponding to it, is particularly acute for searching Hebrew texts. The Responsa project incorporated such morphological analysis of Hebrew words at a very early stage (Choueka and Shapiro 1964; Attar et al. 1978). (For a survey of more recent approaches to this problem and many related problems, see Wintner (2004).)

Second, Hebrew lacks vocalization, so that most words are ambiguous. Thus, tools that exploit context for word disambiguation are of special importance for Hebrew texts. Building on his earlier work on morphological analysis and disambiguation (Choueka and Lusignan 1985), Choueka's work on Nakdan (Choueka and Ne'eman 1995), a tool for automated vocalization, constitutes an implicit form of disambiguation. There has been much recent work on part-of-speech tagging of unvocalized Hebrew text, using both supervised (Bar-Haim et al. 2008) and unsupervised (Adler and Elhadad 2006) disambiguation methods. These models were trained on Modern Hebrew texts and are thus not directly exportable to the Responsa corpus; however, the unsupervised approach should be easily adaptable to the older dialects used in the corpus.

Finally, in Hebrew texts, and especially Rabbinic Hebrew texts, it is not unusual to use abbreviations in the form of acronyms, even for common phrases (and not only named entities as in English). In many of the documents in the Responsa corpus, the proportion of abbreviations to words is about 20% and over one third of them permit more than one expansion (Hacohen-Kerner et al. 2004), thus creating another kind of ambiguity. Recently, progress has been made on disambiguating acronyms generally (Yu et al. 2007) and in Rabbinic Hebrew specifically (Hacohen-Kerner et al. 2008a).

In retrospect, with regard to the linguistic challenges presented by the Responsa corpus, the project was pioneering and met many of the challenges adequately. The implementation of subsequent innovations in disambiguation could further strengthen the project.

3 Searching for Information

As one of the first large corpora, the Responsa corpus required sophisticated search algorithms. Researchers working on the Responsa project designed some of the first efficient algorithms for indexing and compression (Choueka, Fraenkel & Perl 1981, Choueka et al. 1988), for phrase identification (Choueka, Klein & Neuwitz 1983, Choueka 1988) and for proximity-based search.

It is instructive to contrast the proximity-based search implemented in the Responsa project with the well-known vector-space model of Salton (1975). Proximity-based search preserves the full text and returns all documents responsive to a (proximity-dependent) query in unranked form. The vector-space model uses a bag-of-words representation of documents and queries, weights words according to their (inverse) frequency in the corpus, and ranks documents according to cosine distance between a document and a query. Both the preservation and exploitation of word location in a document (an advantage of the Responsa project's method) and the differentiation among varying degrees of responsiveness to a query (an advantage of the vector-space model) are now regarded as crucial to the search for information. The incorporation of both of these properties is now a de facto standard for search as a result of the immense popularity of Internet search engines such as Google.

As noted, a crucial difference between the Responsa corpus and the Brown corpus is that the Responsa corpus is a source of information and not just of language. Thus, a user of the Responsa corpus might typically be searching for a topic, rather than for particular words. One of the well-known problems in topic-based search is that of synonymy. For example, a query for the word *automobile* would not ordinarily return documents that include the word *car* (but not the word *automobile*), even though such a document might be responsive to the user's information need. Many solutions to this problem have been proposed including query expansion using manually constructed thesauri (such as WordNet) or automatically constructed thesauri (Dagan 2000, Lin 1998). The latter might be based on identifying words that have first-degree similarity (that is, they are often collocated) or second-degree similarity (that is, they appear in similar contexts). Other expansion methods, such as automated relevance feedback, can be carried out on the fly: initial results for a query can be examined for words that appear with higher than random frequency that can then be added to the initial query. Some of the earliest work on this method was carried out by Responsa project researchers (Attar and Fraenkel 1977, Hanani 1987). Its implementation would greatly enhance the project's search capabilities.

In Internet search, expansion methods have not yet proved to be as useful as might be expected (in part because they sometimes exacerbate the problem of polysemy –

the phenomenon of single words having multiple meanings). However, the need for such methods is especially acute in the Responsa project. This is because the vast chronological expanse of the corpus renders it especially vulnerable to language drift: the same concepts are often referred to by different terms in different periods. In particular, the modern user might search for concepts using some neologism that, in the best case, would appear only in very recent documents, in the vain hope of finding ancient documents referring to the underlying concept. (Of course, in the extreme case of neologisms that do not appear in the corpus at all, expansion techniques are not helpful unless the corpus is first supplemented with contemporary texts at least some of which include the neologism.)

While this problem might be partially solvable using query expansion, there is another approach for broadening search results that is also promising: exploiting cross-references among documents. Since these are tightly tied to the historical nature of the Responsa corpus, let us first turn to the multitude of issues raised by this historical nature.

4 Chronologically Ordered Corpus

The Responsa corpus spans a period of well over two thousand years and is rich in cross-references. Thus, the Tannaitic literature cites verses from the Bible, the Talmud cites the Tannaitic literature and the Bible, some of the legal codes cite the Talmud, and the responsa cite the Talmud, the legal codes and earlier responsa.

The analysis of such citations has long been used in the bibliometrics community for purposes of document evaluation and information retrieval (Garfield 1972) and its significance for the Responsa project was noted early on by Rabinowitz (1986). In recent years, there has been an explosion of research in this area aimed at exploiting Internet hyperlinks (Brin & Page 1998, Kleinberg 1999) for information retrieval.

We will see below that this work can be leveraged in the Responsa project in a number of ways. First we note that the analogy between citations in the Responsa corpus and Internet hyperlinks is somewhat imperfect for several reasons.

First of all, unlike links, the citations are rarely explicitly marked as such and are generally not characterized by standard forms. Thus, a fundamental challenge is to a)identify a text item as being a citation, b)identify the work that is being referenced and c)identify the specific document being referenced within that work. The design of automated methods for achieving this is a non-trivial task, but one well worth undertaking, as we will see below.

Second, the primary uses of links in a system such as Google are finding documents for the purpose of indexing them and establishing the legitimacy of a document (or site) on the basis of sites linking to it. The Responsa corpus grows in a very controlled manner, so that it can essentially be regarded as a closed set. As a result, citations are not needed for finding documents or for establishing their legitimacy.

Third, since – unlike web documents – corpus documents are static, citations can be used as markers of chronology: a referring document must be subsequent to a document to which it refers.

Bearing in mind all of the above, we can think of the responsa corpus as a directed graph. At a low level of resolution, the nodes of the graph represent authors (or books) and at a high level of resolution the nodes represent documents. In either case, a directed edge from X to Y indicates that X cites Y. More generally, a weighted directed edge reflects the relative frequency of such citations.

These graphs can be exploited in a number of ways that could be useful for the Responsa project.

Beginning with the low-resolution graph, the first thing we observe is that (conflating contemporaries who both cite each other) the graph defines a partial ordering on authors representing the chronological structure of the corpus. Almost all this chronological information is already well known to scholars. But the graph, especially the weighted version, can now be used to precisely measure the flow of information through the generation, to measure the degree of direct and indirect influence one scholar had on another, and to cluster graph nodes into tightly intra-related schools of thought.

Perhaps more importantly, we can use the high-resolution graph to present users with vastly improved results for search queries. We begin by using any standard statistical search method to obtain initial results. We then consider the sub-graph of the high-resolution reference graph that includes only documents included in the initial results. We can then use algorithms similar to PageRank (Brin and Page 1998) or HITS (Kleinberg 1999) to identify among these documents those that are most authoritative or that cite many authoritative documents. We can then use straightforward graph completion techniques to identify relevant documents that were not included in the initial results. Furthermore, we can present results in a (possibly non-linear) manner that reflects the flow of information through the generations and identifies distinct clusters of information flow. These clusters might represent different aspects of a topic (or different senses of a query term) or different schools of thought within the same topic.

We note that the above-mentioned automated techniques can be profitably integrated with manual techniques. For example, instead of initial results being provided by an initial search, users – possibly taking advantage of a platform similar to that of Wikipedia – might provide the central sources on a given topic. The automated methods just described could then be used both to expand the user-generated content and to automatically check it for consistency.

5 Accuracy and Provenance of Texts

Two issues that arise with regard to important historical texts are the accuracy of texts (where variant manuscripts suggest scribal errors or emendations) and provenance of texts (in cases of disputed authorship or unattributed texts). The Responsa project bears a dual relationship with each of these issues: on the one hand, each poses a challenge to the project's ability to maintain corpus quality and, on the other hand, the scope of the project's corpus suggests a number of ways in which it might be used to develop novel methods to address these challenges.

With regard to text accuracy, scholars have developed a variety of essentially heuristic methods for reconciling, or choosing from among, variant manuscripts of the same texts. The availability of electronic versions of these variant manuscripts suggests the possibility of automated processes for reconstructing the most likely original text from these variants. Such processes would necessarily include two main stages. In the first stage, correlations between pairs of manuscripts would be used to establish dependencies between them. In the second stage, manuscripts (or clusters of manuscripts) determined to be pairwise independent could be weighted and aggregated in such manner as to yield maximum likelihood resolutions of disputed text elements. It has been shown in Baharad et al. (2008) how, in the absence of known ground truth for assessing voter (in this case, manuscript) reliability, unsupervised methods, such as EM, can be used to optimally aggregate votes (in this case, readings).

With regard to anonymous texts or cases of disputed authorship, the Responsa corpus can be used to model writing styles of either individual authors or of classes of geographically or chronologically homogeneous classes of authors. Such models can be used to, respectively, identify or profile authors of disputed or anonymous texts. A number of examples of attribution problems that were solved using the Responsa corpus can be found in Mughaz (2003), Koppel et al. (2005) and Hacohen-Kerner et al. (2008b). Thus, for example, it was shown that known responsa of Rashba and Ritba can be used to learn automated classifiers that determine which disputed response were written by each of them. With considerably more difficulty, it was shown that the collection of responsa, *Rav Pe'alim*, for which Rabbi Yosef Haim of Baghdad (*Ben Ish Hai*) acknowledged authorship and the collection, *Torah Lishmah*, for which he denied authorship, were almost certainly written by the same author.

Problems of text accuracy and provenance feature less prominently in the literature surrounding the Responsa project because they are mostly transparent to the project's users. However, proper handling of these issues is ultimately crucial to the user experience. The development and incorporation of tools for ensuring accuracy and correct attribution of texts included in the corpus should play a prominent role in the project in coming years.

6 Conclusions

The Bar-Ilan Responsa project served as a springboard for a good deal of pioneering work on computational linguistics for Hebrew and on foundations of information retrieval. In fact, it is quite astonishing how many ideas that are still at the cutting edge of research in these areas were introduced by Fraenkel and Choueka and their students in the context of the project. Unfortunately, the project’s conversion from a research-oriented undertaking to a commercial enterprise cut short many promising research directions that were fruitfully continued in other venues. The project’s functionality could now be greatly enhanced by implementing a number of techniques some of which were initially proposed and explored in its own laboratory over twenty years ago.

First, the search engine needs to incorporate statistical ranking methods that are now commonplace for all content-based corpora. Second, both manual and automated query-expansion methods need to be introduced, possibly in an interactive manner to prevent potential degradation due to polysemy. Third, cross-references among documents in the corpus need to be (manually or automatically) tagged and exploited in order to present richer and more structured results to users. Fourth, the collective efforts of educated users need to be assembled and organized in Wiki fashion and linked to the corpus. Finally, tools must be developed and incorporated for ensuring accuracy of the texts themselves and of the attribution of texts to specific authors.

The Responsa corpus is already a critical resource for Rabbis, laymen and researchers of Jewish law. In the next few years, however, the project will need to maintain its technological edge if it wishes to remain relevant and to continue to make a contribution to the study of classical Jewish sources.

References

1. Adler, M., Elhadad, M. (2006), An Unsupervised Morpheme-Based HMM for Hebrew Morphological Disambiguation, ACL 2006
2. Attar, R. and Fraenkel, A.S. (1977), Local Feedback in Full-Text Retrieval Systems. J. ACM 24(3), pp. 397-417
3. Attar, R., Choueka, Y., Dershowitz, N. and Fraenkel, A.S. (1978), KEDMA - Linguistic Tools for Retrieval Systems, J. ACM 25(1), pp. 52-66
4. Baharad, E., Goldberger, J., Koppel, M. and Nitzan, S. (2008), Beyond Condorcet: Optimal Judgment Aggregation Using Voting Records, submitted for publication.
5. Bar-Haim, R., Sima'an, K. and Winter, Y. (2008), Part-of-Speech Tagging of Modern Hebrew Text. 2008, Natural Language Engineering 14(2), pp. 223-251
6. Brin, S. and Page, L. (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Networks 30, pp. 107-117
7. Choueka, Y. (1988), Looking for needles in a haystack or: locating interesting expressions in large textual databases, Proc. of the RIAO International Conference on User-Oriented Content-Based Text and Image Handling, pp. 609-623.
8. Choueka, Y. (1990), RESPONSA - A full-text system with linguistic components for large corpora, in Computational Lexicology and Lexicography, a volume in honor of B. Quemada, A. Zampolli (Ed.), Giardini Editions, Pisa, 1990, 181-217.
9. Choueka, Y., Fraenkel, A.S. and Klein, S.T. (1988), Compression of Concordances in Full-Text Retrieval Systems, SIGIR 1988, pp. 597-612
10. Choueka, Y., Fraenkel, A. and Perl, Y. (1981), Polynomial Construction of Optimal Prefix Tables for Text Compression, Proc. of 19th Allerton Conference on Communication, Control and Computing, pp. 762-768
11. Choueka, Y., Klein, S.T. and Neuwitz, E. (1983), Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus, ALLC Journal 4, pp. 34-38
12. Choueka, Y. and Lusignan, S. (1985), Disambiguation by short context, Computers and the Humanities, 19(3), pp. 147-157.
13. Choueka, Y. and Neeman, Y. (1995), Nakdan-Text, Tel-Aviv, C.E.T., 1995.
14. Choueka, Y. and Shapiro, M. (1964), Machine analysis of Hebrew morphology: potentialities and achievements (Hebrew), Leshonenu (Journal of the Academy of Hebrew Language) 27, pp. 354 -372

15. Dagan, I. (2000), Contextual Word Similarity, in Handbook of Natural Language Processing, R. Dale, H. Moisl and H. L. Somers (eds.), CRC Press
16. Garfield, E. (1972), Citation Analysis as a Tool in Journal Evaluation, Science 178(60), pp. 471-479
17. HaCohen-Kerner, Y., Kass, A. and Peretz, A. (2004), Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents. Proc. of the 4th Int'l Conf. on Advances in Natural Language (LNAI), pp. 58-69.
18. Hacohen-Kerner, Y., Kass, A., and Peretz, A. (2008a), Combine One Sense Disambiguation of Abbreviations, Proc. of ACL (Companion Volume), pp. 61-64
19. HaCohen-Kerner, Y., Mughaz, D., Beck, H. and Elchai, Y. (2008b) Words As Classifiers of Documents According to their Historical Period and the Ethnic Origin of their Authors, Cybernetics and Systems,39(3), pp. 213-228.
20. Hanani, S. (1987), Feedback by Local Clustering in a Full-text Online Information Retrieval System, Unpublished M.Sc. Thesis, Bar-Ilan Iniversity, 1987.
21. Kleinberg, J. (1999), Authoritative sources in a hyperlinked environment, Journal of the ACM 46 (5), pp. 604–632
22. Koppel, M., Mughaz, D. and Akiva, N. (2006), New Methods for Attribution of Rabbinic Literature , Hebrew Linguistics: A Journal for Hebrew Descriptive, Computational and Applied Linguistics 57, pp. 5-18.
23. Kucera, H., and Francis, W.N. (1967), Computational Analysis of Present-day American Engish, Providence: Brown University Press
24. Lin, D. (1998), Automatic Retrieval and Clustering of Similar Words, COLING-ACL 1998, pp. 768-774.
25. Mughaz, D. (2003). Classification of Hebrew texts according to style. M.Sc. thesis (in Hebrew), Bar-Ilan University, Ramat-Gan, Israel.
26. Rabinowitz, R. (1986), Performance Improvement of the Information Retrieval Systems Based on Utilization of the References Included in the Retrieved Documents, Unpublished M.Sc. Thesis, Bar-Ilan Iniversity, 1986.
27. Salton, G., Wong, A. and Yang, C.S. (1975), A Vector Space Model for Automatic Indexing, Commun. ACM 18(11), pp. 613-620
28. Wintner, S. (2004), Hebrew computational linguistics: Past and future. Artificial Intelligence Review, 21(2):113-138
29. Yu, H., Kim, W., Hatzivassiloglou, V., Wilbur, W.J. (2007), Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles, Journal of Biomedical Informatics 40(2), pp. 150-159.