Estimating the Birth and Death Years of Authors of Undated Documents Using Undated Citations

Yaakov HaCohen-Kerner[1] and Dror Mughaz[2,1]

[1] Dept. of Computer Science, Jerusalem College of Technology, 91160 Jerusalem, Israel
[2] Dept. of Computer Science, Bar-Ilan University, 52900 Ramat-Gan, Israel
kerner@jct.ac.il, myghaz@cs.biu.ac.il

Abstract. Precious historical treasures might be hidden between the lines of a text. There are many implicit details which can be extracted from a text, particularly if one has access to an entire corpus of texts pertaining to the given subject. One of these details is the identification of the era in which the author of the given document(s) lived. For rabbinic documents written in Hebrew and Aramaic, which are almost without exception undated and do not contain any bibliographic section, this problem is extremely important. The aim of this novel research is to find in which years an author was born and died, based on his documents and the documents of other authors (whose birth and death years are known) who refer to the author under discussion or are mentioned by him. Such estimates can help determine the time frame in which certain documents were written and in some cases identify an anonymous author. In the framework of this research, we formulate various kinds of "iron-clad", heuristic and greedy constraints defining the birth and death years of an author based on citations referring to him or mentioned by him. Experiments applied on a corpus containing texts composed by rabbinic authors show reasonable results.

Keywords: Citation analysis, Hebrew, Hebrew-Aramaic documents, knowledge discovery, time analysis, undated citations, undated documents.

1 Introduction

Citations are a defining feature of many kinds of documents, e.g., academic, legal and religious. Authors cite previous works which are related in some way to their own work or to their discussion. Citations included in documents are important information resources of interest to researchers. Therefore, automatic extraction and analysis of citations from documents are of great importance.

Recent developments (e.g., computerized corpora and search engines) enable accurate extraction of citations. As a result, citation analysis has an increased importance.

A citation is a brief reference in the body of the text to a source of published information. A reference includes bibliographic details about a source that is mentioned in a citation. The reference is found at end of a document in a reference list. Citations are presented in agreed typographical formats. Different disciplines have different conventions: citation in footnotes, citations with numbers (e.g., [1]) or mixed symbols such as [Cohen98] or [Cohen 1998] (Harvard-style citations).

H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (Eds.): IceTAL 2010, LNAI 6233, pp. 138–149, 2010.

Garfield [2] was the first to propose automatic production of citation indexes, extraction and analysis of citations from corpora of academic papers. Powley and Dale [5] develop techniques to extract from a given academic paper a list of citations and, for each citation, the corresponding reference in the reference list. They find each instance of a citation in the body of the paper; parse it into a set of author names and years; and find the segment of text from the references which contains the corresponding reference.

Teufel et al. [8] use extracted citations and their context for automatic classification of citations to their citation function (the author's reason for citing a given paper). Some research has been done concerning the improvement of retrieval performance using terms. Ritchie et al. [6] show that document indexing based on combinations of terms used by citing documents and terms from the document itself give better retrieval performance than standard indexing of the document terms alone. In [7], Ritchie et al. investigate how to select text from around the citations in order to extract good index terms in order to improve retrieval effectiveness.

Citations are a defining feature not just of academic papers but also and even more of rabbinic responsa (answers written in response to Jewish legal questions authored by rabbinic scholars). Citations included in rabbinic literature are more complex to define and to extract than citations in academic papers written in English because:

(1) In contrast to academic papers, there is no reference list that appears at the end of a responsa;

(2) There is an interaction with the complex morphology of Hebrew and Aramaic. For example, citations can be presented with different types of prefixes (e.g., "and …", "when …", "and when …", "in …", "and in …", "and when in …") included in the citation-word(s);

(3) Natural language processing in Hebrew and Aramaic has been relatively little studied;

(4) Many citations in Hebrew-Aramaic documents are ambiguous. For instance: (a) a book titled מגן-אבות *magen-avot* was composed by four different Jewish authors; and (b) The abbreviation מ"ב (m"b) relates to two different Jewish authors and has also other meanings, which are not authors' names; and

(5) At least 30 different syntactic styles are used to present citations. This number is higher than the number of citation patterns used in academic papers written in English (e.g., see [5]).

Each specific document written by a specific author can be referred to, in at least 30 general possible citation syntactic styles. Furthermore, each citation pattern can be expanded to many other specific citations by replacing the name of the author and/or his book/responsa by each one of their other names (e.g., different spellings, full names, short names, first names, surnames, and nicknames with/without title) and abbreviations.

The citation recognition in this research is done by comparing each word to a list of 298 known authors and many of their books/responsa. This list contains 19,506 specific citations that relate to names, nick names and abbreviations of these authors and their writings. Basic known citations were collected and all other citations were produced from them, based on an automatic extension process using regular expressions.

Hebrew-Aramaic documents in general and Hebrew-Aramaic responsa in principle present various interesting text mining problems. Firstly, Hebrew is richer in its morphology forms than English. According to linguistic estimates, Hebrew has 70,000,000 valid (inflected) forms while English has only 1,000,000 [1]. In Hebrew, there are up to seven thousand declensions for one stem, while in English there are only a few declensions. Secondly, these kinds of documents include a high rate of abbreviations (about 20%), while more than one third of them (about 8%) are ambiguous [4].

A previous research that works on corpora, which contain responsa referring to Jewish law written in Hebrew-Aramaic dealt with text classification [3]. In this research, HaCohen-Kerner et al. investigate whether the use of stylistic feature sets and/or name-based feature sets is appropriate for classification of documents to the ethnic group of their authors and/or periods of time when the documents were written and/or places where the documents were written. In addition, HaCohen-Kerner et al. [4] have experience with the processing of such texts from the viewpoint of disambiguation of ambiguous abbreviations. The current research is a continuation of this long-term research interest.

In this research, we present a novel model that estimates the birth and death years of a given author using undated citations of other authors (whose birth and death years are known) who refer to him or mentioned by him. The documents are undated (non-time-stamped) and mentions of years or historical events in the documents are very rare. The estimations are based on various constraints of different degree of certainty: "iron-clad", heuristic and greedy constraints. The constraints are based on general citations without cue words and citations with cue words, such as father, son, rabbi, teacher, student, friend, and "late" ("of blessed memory").

This paper is organized as follows: Section 2 presents various constraints of different degree of certainty: "iron-clad", heuristic and greedy constraints that are used to estimate the birth and death years of responsa authors. Section 3 describes the model. Section 4 introduces the tested dataset, the results of the experiments and their analysis. Section 5 summarizes, concludes and proposes future directions.

2 Citation-Based Constraints

This section presents the citation-based constraints formulated for the estimation of the birth and death years of an author X based on his documents and on other authors' (Y_i) documents who mention X or one of his documents. We assume that the death years (for those who died) and birth years of all authors are known, excluding those of the investigated author. Below are given some notions and constants that are used: X – The author under consideration, Y_i – Other authors, B – Birth year, D – Death year, MIN – Minimal age (currently 30 years) of a rabbinic author when he starts to write his response, MAX – Maximal life period (currently 100 years) of a rabbinic author, and MIN_FATHER – Minimal age (currently 20 years) of a rabbinic author when his firstborn son is born.

The estimations of MIN, MAX and MIN_FATHER constants are only heuristic, although they are realistic on the basis of typical responsa authors' lifestyle.

Various types of citations exist: general citations without cue words and citations with cue words, such as: father, son, rabbi, teacher, student, friend, and "late" ("of blessed memory"). Another classification of the discussed citations is to those referring to living authors and those referring to dead authors. In contrast to academic papers, responsa include much more citations to dead authors than to living authors.

We will introduce citation-based constraints of different degrees of certainty: "iron-clad" (I), heuristic (H) and greedy (G). "Iron-clad" constraints are absolutely true, without any exception. Heuristic constraints are almost always true. Exceptions can occur when the heuristic estimates for MIN, MAX and MIN_FATHER are incorrect. Greedy constraints are rather reasonable constraints for responsa authors. However, sometimes wrong estimates can be drawn while using these constraints. Each constraint will be numbered and its degree of certainty will be presented in brackets.

2.1 "Iron-clad" and Heuristic Constraints

First of all, we present two general heuristic constraints based on authors that cite X, which are based on regular citations (i.e., without mentioning special cue words, e.g., friend, son, father and rabbi).

General constraint based on authors that were cited by X

$$D(X) >= MAX(B(Yi))+MIN \quad (1\ (H))$$

X must be alive when he cited Y_i, so we can use the earliest possible age of publishing of the latest born author Y as a lower estimate for X's death year.

General constraint based on authors that cite X

$$B(X) <= MIN(D(Yi))-MIN \quad (2\ (H))$$

All Y_i must have been alive when they cited X, and X must have been old enough to publish. Therefore, we can use the earliest death year amongst such authors Y_i as an upper estimate of X's earliest possible publication age (and thus his birth year).

Posthumous citation constraints

Posthumous constraints estimate the birth and death years of an author X based on citations of authors who refer to X as "late" ("of blessed memory") or on citations of X who mentions other authors as "late". Figure 1 describes possible situations where various kinds of authors Y_i (i=1, 2, 3) refer to X as "late". The lines depict authors' life spans where the left edges represent the birth years and the right edges represent death years. In this case (as all Y_i refer to X as "late"), we know that all Y_i died after X (and some of the Y_i might be still alive), but we do not know when they were born in relation to X's birth. Y_1 was born before X's birth; Y_2 was born after X's birth but before X's death; and Y_3 was born after X's death.

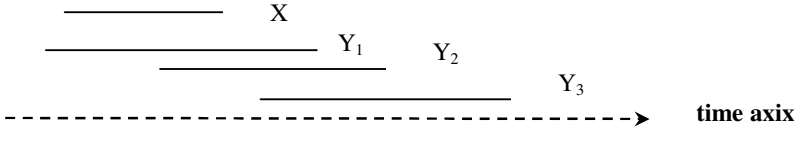

Fig. 1. Citations mentioning X as "late"

$$D(X) <= MIN(D(Yi)) \quad (3\ (I))$$

However, we know that X must have been dead when Y_i cited him as "late", so we can use the earliest born such Y's death year as an upper estimate for X's death year. Like all authors, dead authors of course have to comply to constraint (2) as well.

Let us now look at the cases where the author X, we are studying refers to other authors Yi as "late". Figure 2 describes possible situations where X refers to various kinds of authors Y_i (i = 1, 2, 3) as "late". All Y_i died before X's death (or maybe X is still alive). Y_1 died before X's birth; Y_2 was born before X's birth and died when X was still alive; and Y_3 was born after X's birth and died when X was still alive.

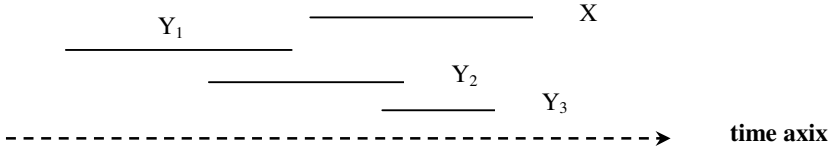

Fig. 2. Citations by X who mentions others as "late"

$$D(X) >= MAX(D(Yi)) \quad (4\ (I))$$

X must be alive after the death of all Y_i who were cited as "late" by him. Therefore, we can use the death year of the latest-born such Y as a lower estimate for X's death year.

$$B(X) >= MAX(D(Yi)) - MAX \quad (5\ (H))$$

X was probably born after the death year of the latest-dying person, who X wrote about. Therefore, we can use the death year of the latest-born such Y minus his maximal life-period as a lower estimate for X's born year.

Contemporary citation constraints

Contemporary citation constraints calculate the upper and lower bounds of the birth year of an author X based only on citations of known authors who refer to X as their friend/student/rabbi. This means there must have been at least some period in time when both were alive and intellectually active. Figure 3 describes possible situations where various kinds of authors Y_i refer to X as their friend/student/rabbi. Y_1 was born before X's birth and died before X's death; Y_2 was born before X's birth and died after X's death; Y_3 was born after X's birth and died before X's death; and Y_4 was born after X's birth and died after X's death. Like all authors, contemporary authors of course have to comply to constraints 1 and 2 as well.

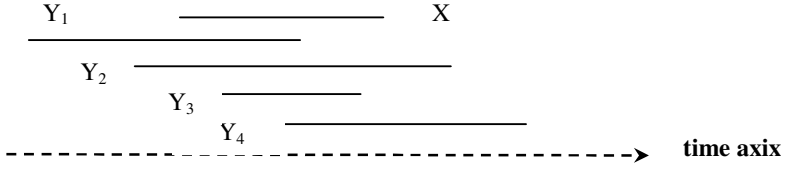

Fig. 3. Citations by authors who refer to X as their Friend/Student/Rabbi

$$B(X)>= MIN(B(Yi))-(MAX-MIN) \quad (6\ (H))$$

All Y_i must have been alive when X was alive, and all of them must have been old enough to publish. Therefore, X could not be born MAX-MIN years before the earliest birth year amongst all authors Y_i.

$$D(X)<=MAX(D(Yi))+(MAX-MIN) \quad (7\ (H))$$

Again, all Y_i must have been alive when X was alive, and all of them must have been old enough to publish. Thus, X could not be alive MAX-MIN years after the latest death year amongst all authors Y_i.

Intellectual son/father-based constraints

Son-based constraints calculate the upper and lower bounds of the birth and death years of an author X based only on citations of only one known author who refers to X as his son. According to rabbinic conventions, X can be either a "truly son" (i.e., a biological son), or an "intellectual son" (i.e., a student).

Figure 4 describes five possible situations. Y_i (i = 1, 2, 3) refer to X as their "truly son". In all these cases, Y_i were born before X's birth. Y_1 died before X's birth (maximum 9 months before X's birth); Y_2 died before X's death; and Y_3 died after X's death. Y_1 is not a possible father in the discussed context, since in this case, Y_1 cannot refer to his son who was born only after Y1's death. However, in Jewish rabbinic documents, it is possible that an author Y_i (e.g. Y_4 or Y_5) will call his student X, a son (meaning an intellectual son), although X is not his "truly son". In such a case, Y_i (the "father") can be born even after X's birth.

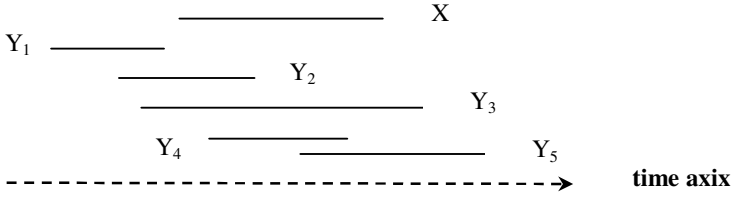

Fig. 4. Citations by authors who refer to X as their son

When taking into account situations such as an intellectual son (X towards Y_4 or Y_5), all son-based constraints are expressed by the friend/student/rabbi-based constraints (6-7). If a biological bond, i.e., a "truly son" can be absolutely identified, than a unique constraint can be formulated.

Father-based constraints calculate the upper and lower bounds of the birth and death years of an author X based only on citations of known authors who refer to X as their father. Also here, according to rabbinic conventions, X can be either a "truly father" (i.e., a biological father), or an "intellectual father" (i.e., a rabbi or a teacher). Therefore, all father-based constraints are expressed by the friend/student/rabbi-based constraints (6-7).

2.2 Greedy Constraints

We also formulate and apply greedy constraints. These bounds are sensible in many cases, but which can nevertheless sometimes lead to wrong estimates. It is important to mention that the greedy constraints are applied in combination with the iron-clad and heuristic constraints. This is because, in many cases some of the greedy constraints are not applied because lack of explicit citations (citations with cue words). In such cases, we use the estimations that are products of the iron-clad and heuristic constraints.

Greedy constraint based on authors who are mentioned by X

$$B(X) >= MAX(B(Yi)) \quad (8\ (G))$$

Most of the citations in our research domain, relate to dead authors. Thus, most of the citations mentioned by X relate to dead authors. That is, most of Y_i were born before X's birth and died before X's death. Therefore, a greedy assumption will be that X was born no earlier than the birth of latest author mentioned by X.

Greedy constraint based on authors who refer to X

$$D(X) <= MIN(D(Yi)) \quad (9\ (G))$$

As mentioned above, most of the citations mentioned by Y_i relate to X as dead. Therefore, most of Y_i die after X's death. Therefore, a greedy assumption will be that X died no later than the death of the earliest author who refers to X.

Refinement of constraints (8-9) are presented by constraints (10- 13). Constraints (10-11) are due to X citing Y_i and Constraints 12-13 are due to Y_i citing X.

Greedy constraint for defining the birth year based only on authors who were cited by X

$$B(X) >= MAX(D(Yi)) \quad (10\ (G))$$

When taking into account only citations that are cited by X, most of the citations, relate to dead authors. That is, most of Y_i died before X's birth. Therefore, a greedy assumption will be that X was born no earlier than the death of the latest author mentioned by X.

Greedy constraint for defining the birth year based only on authors who are mentioned by X as a friend

$$B(X) <= MIN(B(Yi)) \quad (11\ (G))$$

When taking into account only citations that are mentioned by X, which relate to contemporary authors, a greedy constraint can be that X was born no later than the birth of the earliest author mentioned by X as a friend.

Greedy constraint for defining the death year of X based only on authors who cited X as "late"

$$D(X) <= MIN(B(Yi)) \quad (12\ (G))$$

When taking into account only citations that are mentioned by Y_i who relate to X as "late", a greedy assumption can be that X died no later than the birth of the earliest author who cited X as "late".

Greedy constraint for defining the death year of X based only on authors who cited X as a friend

$$D(X) >= MAX(D(Yi)) \quad (13\ (G))$$

When taking into account only citations that are mentioned by Yi who cited X as a friend, all Y_i must have been alive when X was alive, and all of them must have been old enough to publish. Therefore, a greedy assumption will be that X died no earlier than the death of the latest author who cited X as a friend.

We do not present greedy constraints regarding son and father because they can be intellectual son and father and not truly relatives.

3 The Model

The main steps of the model are presented below. Most of these steps were processed automatically, except for steps 2 and 3 that were processed semi-automatically.

1. **Cleaning the texts.** Since the responsa may have undergone some editing, we must make sure to ignore possible effects of differences in the texts resulting from variant editing practices. Therefore, we eliminate all orthographic variations.
2. **Normalizing the citations in the texts.** For each author, we normalize all kinds of citations that refer to him (e.g., various variants and spellings of his name, books, documents and their nicknames and abbreviations). For each author, we collect all citation syntactic styles referred to him and then replace them to a unique string.
3. **Building indexes**, e.g., authors, citations to "late"/friend/student/rabbis/son/father and calculating the frequencies of each item.
4. **Citation identification** into various categories of citations, including self-citations.
5. **Performing various combinations of "iron-clad" and heuristic constraints** on the one hand, **and greedy constraints** on the other hand, **to estimate** the birth and death years for each tested author.
6. **Calculating averages and std-deviations** for the best "iron-clad" and heuristic version and the best greedy version.

4 Experimental Results

The examined dataset includes 3,488 responsa[1] authored by 12 Jewish rabbinic scholars, two of whom are still alive. All these authors lived in the last 130 years and were very productive regarding the number of documents and citations that were written by them. On average, there are about 291 documents for each scholar. These responsa were written in the last 100 years. The total number of words is about 6,887,351 words (average per documents is 1,975 words). This corpus includes citations to 298

[1] Contained in the Global Jewish Database (The Responsa Project at Bar-Ilan University). Http://www.biu.ac.il/ICJI/Responsa

authors including the 12 investigated authors. The dataset before the normalization step (step # 2 in section 4) includes 106,923 citations (i.e., mentions of other works), which are about 8,910 citations in average for each author and about 31 citations for each document. 19,506 of these citations are different.

Since this dataset represents a special corpus containing responsa authored by 12 authors who lived in the last 130 years, the incoming posthumous citations count is always 0. This special situation enables us to correct death ages which are higher than the current year. That is, if the upper bound of D(X) is greater than the current year then we change it to the current year. If the investigated authors died a few hundreds years ago, then the upper bounds would probably been much worse.

The situation with these authors also means that we did not apply average posthumous constraints (greedy rules # 8, 10 for the birth year and greedy rules # 9, 12 for the death year). In a different corpus situation (where all authors are roughly from the same period), these greedy rules help, but not here, where many ancient authors are cited (i.e., some of the lower bounds can be hundreds years ago and if we use them than the estimation for B(X) will be too low and therefore very bad).

Several characteristics of this dataset are presented below:

On average, each author cites 8,910 citations while only about 10 of them are posthumous citations and about 6 of them are contemporary citations. About 99.8% of the citations are implicit, i.e., they are not accompanied with cue words that identify whether the citations are posthumous or contemporary.

The average number of citations to each author is 88 including self citations and 33 excluding self citations. That is, most of the citations (62.5%) are self citations.

Among the explicit citations (those with cue-words) the average number of posthumous citations (10.25) is about twice greater than the average number of contemporary citations (5.67). That is, about two-thirds of the explicit citations are posthumous.

On average, for each author there are much more outgoing citations (8,910) than incoming citations (88) in general and more outgoing contemporary citations (6) than incoming contemporary citations (4).

Table 1 compares the ground truth about the birth and death years on the one hand to the best iron-clad and heuristic version and on the other hand to the best greedy version.

Since this is a novel problem, it is difficult to evaluate the results in the sense that although we can compare how close the system guess is to the actual birth/death years, what we cannot do is assess how-close-is-close, i.e. there is no real notion of what a 'good' result is.

Currently, we use the notion difference, which is defined as the estimated value minus the ground truth value. Some of the estimates for birth and death years are not integer values. This finding is due to the use of average functions in certain versions (e.g., two last sub-rows in tables 2 and 3).

Table 1 shows that the best experimental results have been achieved by the best greedy version, which was better than the best iron-clad and heuristic version as follows: (1) Its average birth-year and death-year differences (13.04 and 15.54, respectively) are better than those of the best iron-clad and heuristic version (22 and 22.67, respectively), (2) The absolute differences of 12 out of 24 estimates were less or equal to 6.5 years, versus only 5 such estimates of the best iron-clad and heuristic version and (3) The standard deviation of the birth-year's greedy estimate is less than its comparable

iron-clad and heuristic standard deviation. This indicates that the results of the best greedy version are steadier.

Indeed, the best greedy version was better than the best iron-clad and heuristic version only in 14 out of 24 estimates (of birth and death years). Therefore, these results are still not enough significant.

Table 2 presents the experimental results using the various iron-clad and heuristic constraints only (section 2.1). The minimal average birth-year and death-year differences (22 and 22.67, respectively) have been achieved by the version of the average "late"-based constraints (constraints 3-6). This result was obtained using the average

Table 1. Experimental results using various groups of constraints

	Author X	Ground truth		Best iron-clad & heuristic version		Differences for best iron-clad & heuristic version		Best greedy version		Differences for best greedy version	
#	Name of X (in Hebrew)	Birth year	Death year	Birth year	Death year	Birth year	Death year	Birth year	Death year	Birth year	Death year
1	אליעזר וולדינברג	1917	2006	1879	1971.5	38	34.5	1899.5	1953	17.5	53
2	בצלאל שטרן	1911	1989	1885	1959.5	26	29.5	1910	1989	**1**	**0**
3	עובדיה יוסף	1920	Alive	1888.5	1981	31.5	29	1894	1953	26	57
4	בן-ציון עוזיאל	1880	1953	1862.5	1952	17.5	**1**	1884	1959.5	**-4**	**-6.5**
5	יצחק הרצוג	1889	1959	1888.5	1981	**0.5**	-22	1874.5	1958	14.5	**1**
6	יצחק וויס	1902	1989	1887	1958.5	15	30.5	1880.5	1995	21.5	**-6**
7	יעקב עדס	1898	1963	1857.5	1950	40.5	13	1885	1980.5	13	-17.5
8	משה פיינשטיין	1895	1986	1913.5	1988	-18.5	**-2**	1889	1959	**6**	27
9	עובדיה הדאיה	1890	1969	1833.5	1923	56.5	46.5	1889	1971.5	**1**	**-2.5**
10	רחמים חוויתה	1901	1959	1915.5	1980.5	-14.5	-21.5	1874.5	1950	26.5	9
11	שמואלווזנר	1914	Alive	1916	1981	**-2**	29	1920	2009	**-6**	**1**
12	שלמה זלמן אויערבך	1910	1995	1906.5	1981	**3.5**	14	1890.5	1989	19.5	**6**
					Ave.	**22**	**22.67**		**Ave.**	**13.04**	**15.54**
					Std. dev.	17.15	13.28		Std. dev.	9.32	20.00

Table 2. Experimental results using different groups of constraints

Group of cons.	Upper and lower bounds	Average of absolute differences (in years)	
		Birth year	Death year
Cons. 1-2	B(X)< , D(X)>	35.83	38.67
Posthumous citation cons. (cons. 2-3)	B(X)< , D(X)<	43.42	26.33
	B(X)< , D(X)>	43.42	55.83
	B(X)> , D(X)<	43.42	26.33
	B(X)> , D(X)>	75.75	55.83
(cons. 2,5) & (cons. 3,4)	B(X)=ave(B(X)<, B(X)>)	22.00	
	D(X)=ave(D(X)<, D(X)>)		22.67
Contemporary cons. (cons. 1-2, 4-5)	B(X)< , D(X)<	37.58	33.67
	B(X)< , D(X)>	37.58	38.25
	B(X)> , D(X)<	87.58	33.67
	B(X)> , D(X)>	87.58	38.25
	B(X)=ave(B(X)<, B(X)>)	45.08	29.79
	D(X)=ave(D(X)<, D(X)>)		

of the upper and the lower bounds of the birth year as estimate for the birth year and the average of the upper and the lower bounds of the death year as estimate for the death year. This version is better than the version that contains the two most simple constraints (1-2), which do not take into consideration any cue-words. This finding indicates that the posthumous and contemporary constraints do contribute to the estimates.

The result achieved by the best iron-clad version was successful also because an important correction that was done by us concerning the iron-clad constraints dealing with the estimation of D(X). That is, if the upper bound of D(X) is greater than the current year then we change it to the current year. If the investigated authors died a few hundreds years ago, then the upper bounds would probably been much worse. In general, the results achieved by the contemporary (friend) constraints were worse than those achieved by the "late" constraints. That might be due to the fact that there more posthumous citations than contemporary citations.

Table 3. Experimental results using different versions of the greedy constraints

Group of cons.	Average of absolute differences (in years)	
	Birth year	Death year
Cons. 8-9	13.42	17.30
Posthumous cons. (10, 12)	43.42	26.30
Contemporary cons. (1-2, 4-5)	37.58	33.67
Ave. friend cons. B(X) = ave(10,11), D(X) = ave(12,13)	13.04	15.54

Table 3 presents the results achieved by the different versions of the greedy constraints (section 2.2). The minimal averages of absolute differences (in years) for the birth and death years (13.04 and 15.54, respectively) have been achieved by the greedy version of the average "friend"-based constraints (constraints 10-13).

5 Summary, Conclusions and Future Work

To the best of our knowledge, we are the first to investigate the estimation of the birth and death years of the authors using undated citations referring to them or written by them. This investigation was performed on a special case of documents (i.e., responsa), where special writing rules are applied. The estimation was based on the author's documents and documents of other authors (whose birth and death years are known) who refer to the discussed author or are mentioned by him. To do so, we formulate various kinds of iron-clad, heuristic and greedy constraints. The best estimates have been achieved using the version of the average contemporary greedy constraints.

Regarding the estimation of the birth and death years of an author X, it is important to point that citations mentioned by X or referring to X are more suitable to estimate the "birth" and "death" writing years of X rather than his real birth and death years.

This model might be applied with suitable changes to similar research problems that might be relevant for some historical legal or religious document collections. Usually, such documents include citations to previous documents of the same kind.

We plan to improve the estimation of the birth and death years of authors by: (1) Combining and testing new combinations of iron-clad, heuristic and greedy constraints, (2) Improving existing constraints and/or formulating new constraints (e.g., statistical-based constraints), (3) Defining and applying heuristic constraints that take into account various details included in the responsa, e.g., dates (in case that they appear), events, names of people, concepts, special words and collocations that can be dated, (4) Conducting additional experiments using many more responsa written by more authors is supposed to improve the estimates, (5) Checking why the iron-clad, heuristic and greedy constraints tend to produce more positive differences, and (6) Testing how much of an improvement we got from a correction of the upper bound of D(x) and how much we will at some point use it for a corpus with long-dead authors.

Definition and application of additional kinds of constraints is planned: (1) Constraints that are based on historical events mentioned in the documents; and (2) Three-generation constraints, i.e., constraints that relate to biological or preceding relations, e.g., grand son and grand student. Another interesting future research is the disambiguation of ambiguous citations.

Acknowledgements. The authors thank Simone Teufel for reviewing drafts of this article and offering many helpful comments, and three anonymous reviewers for their reviews.

References

1. Choueka, Y., Conley, E.S., Dagan, I.: A Comprehensive Bilingual Word Alignment System: Application to Disparate Languages - Hebrew, English. In: Veronis, J. (ed.) Parallel Text Processing, pp. 69–96. Kluwer Academic Publishers, Dordrecht (2000)
2. Garfield, E.: Can Citation Indexing be Automated? In: Stevens, M. (ed.) Statistical Association Methods for Mechanical Documentation, Symposium Proceedings, vol. 269, pp. 189–142. National Bureau of Standards Miscellaneous Publication (1965)
3. HaCohen-Kerner, Y., Beck, H., Yehudai, E., Rosenstein, M., Mughaz, D.: Cuisine: Classification using Stylistic Feature Sets and/or Name-Based Feature Sets. To appear in Journal of the American Society for Information Science and Technology, JASIST (2010) (Published Online: Apr 22 2010), (DOI: 10.1002/asi.21350)
4. HaCohen-Kerner, Y., Kass, A., Peretz, A.: HAADS: A Hebrew Aramaic Abbreviation Disambiguation System. To appear in Journal of the American Society for Information Science and Technology, JASIST (2010) (Published Online: May 27, 2010), (DOI: 10.1002/asi.21367)
5. Powley, B., Dale, R.: Evidence-Based Information Extraction for High Accuracy Citation and Author Name Identification. In: RIAO 2007 (2007)
6. Ritchie, A., Teufel, S., Robertson, S.: Using Terms from Citations for IR: Some First Results. In: The European Conference for Information Retrieval (ECIR), pp. 211–221 (2007)
7. Ritchie, A., Robertson, S., Teufel, S.: Comparing Citation Contexts for Information Retrieval. In: The 17th ACM Conference on Information and Knowledge Management (CIKM), pp. 213–222 (2008)
8. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic Classification of Citation Function. In: The 2006 Conference on Empirical Methods in Natural Language Processing, ACL, pp. 103–110 (2006)