AUTOMATIC IDENTIFICATION OF BIBLICAL QUOTATIONS IN HEBREW-ARAMAIC DOCUMENTS

Yaakov HaCohen-Kerner[1] , Nadav Schweitzer[2] and Yaakov Shoham[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*

Keywords: Hebrew-Aramaic Texts, Information Retrieval, Quotation identification.

Abstract: Quotations in a text document contain important information about the content, the context, the sources that the author uses, their importance and impact. Therefore, automatic identification of quotations from documents is an important task. Quotations included in rabbinic literature are difficult to identify and to extract for various reasons. The aim of this research is to automatically identify Biblical quotations included in rabbinic documents written in Hebrew-Aramaic. We deal with various kinds of quotations: partial, missing and incorrect. We formulate nineteen features to identify these quotations. These features were divided into seven different feature sets: matches, best matches, sums of weights, weighted averages, weighted medians, common words, and quotation indicators. Several features are novel. Experiments on various combinations of these features were performed using four common machine learning methods. A combination of 17 features using J48 (an improved version of C4.5) achieves an accuracy of 91.2%, which is an improvement of about 8% compared to a baseline result.

1 INTRODUCTION

A quotation is a repetition of either an expression or a sub-expression as part of another one. The quoted expression is well-known (e.g., a Biblical quotation) or explicitly attributed to its original source (using a citation).

Recent developments (e.g., computerized corpora and search engines) enable massive and accurate extraction of quotations. As a result, quotation analysis has attained increased importance.

There are various reasons for using quotations: (1) to demonstrate or to clarify or to reinforce the claims of the work that uses the quotation, (2) to supply information about the quoted work, (3) to respect the quoted work or author, and (4) to obey to the copyright law.

Many quotations are taken from the Bible. Other usable quotations can be found in quotation dictionaries such as: The Oxford Dictionary of Quotations, The Yale Book of Quotations, the Bartlett's Familiar Quotations, and The MacMillan Book of Proverbs, Maxims and Famous Phrases.

In many cases, there are various kinds of problems concerning quoting quotations, e.g., (1) quotations are incorrect and/or (2) quotations are partial (e.g., words are missing) and/or (3) quotations are attributed to wrong authors.

Quotations are a defining feature of many kinds of documents, e.g., religious, legal and academic. Quotation analysis can supply important information about the content, the context, the sources that the author uses, their importance and impact.

Quotations are also an important and common feature of rabbinic responsa (answers written in response to Jewish legal questions). Quotations included in rabbinic literature are more difficult and challenging to identify and to extract than quotations in academic papers written in English because:

(1) While academic papers typically employ one of a small number of standard styles for quotations, rabbinic quotations are essentially free-form. Each quotation can be written in many possible styles;

(2) Academic quotations are typically attributed (e.g., by a reference or a footnote), while a Hebrew-Aramaic quotation, in many cases, is not attributed;

(3) There is an interaction with the complex morphology of Hebrew and Aramaic, which have a richer system of inflectional morphology than English (Choueka et al., 2000). For instance, quotations can be presented with different types of

Hacohen-Kerner Y., Schweitzer N. and Shoham Y..
AUTOMATIC IDENTIFICATION OF BIBLICAL QUOTATIONS IN HEBREW-ARAMAIC DOCUMENTS.
DOI: 10.5220/0003106703200325
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 320-325
ISBN: 978-989-8425-28-7

prefixes included in the name of a citation.

There have been various researches on quotations in information retrieval (IR). However, our research is unique in addressing the much more difficult problem of identifying quotations included in rabbinic documents written in Hebrew-Aramaic. We defined and applied various features for the identification process. Some of them are novel.

Given both the Bible and texts that include quote parts of the Bible, we aim to identify various kinds of quotations: partial, missing and incorrect.

Currently, we deal only with direct Biblical quotations that are presented in various styles. More complex cases, e.g., nested quotations and quotations attributed to unspecified authors are left for future research.

The identification of quotations is based on various feature sets used in conjunction with four common machine learning (ML) methods.

This paper is organized as follows: Section 2 gives background concerning quotation extraction. Section 3 presents different styles used to present quotations. Section 4 describes the feature sets built for the identification of quotations in Hebrew-Aramaic documents. Section 5 presents the model. Section 6 introduces the examined dataset, the results of the experiments and their analysis. Section 7 summarizes and proposes future directions.

2 QUOTATION EXTRACTION

Liang et al. (2010) developed a system that automatically extracts quotations from news feeds, and allows efficient retrieval of the semantically annotated quotes. For each text document, the system splits the document into paragraphs, the paragraphs into sentences, and parse each sentence using linguistic parsing, lexical analysis and determining grammatical roles and named entity recognition. They apply co-reference resolution to identify multiple mentions of the same entity across the whole document, including resolving pronoun co-reference, aliases and abbreviations. For the underlying information retrieval capability, they use a typical Vector SpaceModel (VSM) based system. They enable querying over 10 million quotes extracted from news feeds. In addition, each day they add about 60,000 new quotes. Based on a test set of 150,000 randomly sampled entities, the uncached query execution time has an average of 109 milliseconds with median at 54 milliseconds.

Pouliquen et al. (2007) built a system capable to automatically identify direct speech quotations in texts written in 11 languages. Their system identifies the quotations and the people who made the quotations, and where applicable, the people or organizations mentioned in the quotations. The system also identifies variants of each name. They rely mainly on lexical patterns with character level regular expressions, which are easily transposable to new languages.

To detect quotations, they use a simple method that search for quotation markers found close to reporting verbs (say, tell, declare, etc.) and known person names. They identify quotations using a small number of rules: (1) quotation marker identification (quote-characters like “, ”, «, » etc.), (2) reporting verbs (e.g., confirmed, says, declared), (3) general modifiers, which can appear close to the verb (e.g., the adverb yesterday), (4) determiners, which can appear between the verb and the person name (e.g., a, an, the), (5) trigger-for-person (e.g., British Prime Minister), (6) person name (e.g., Tony Blair), and (7) a list of matching rules (e.g., name verb [adverb] quote-mark QUOTE quote-mark)

Using this process they gather an average of 2,665 quotes per day. As of June 2007, they have a repository of about 1,500,000 quotes, gathered during two years of analysis.

To evaluate the recall of the quotation recognition system, they searched a random collection of news articles (documents dated 12 July 2007) and carried out a manual evaluation for 55 of the quotations found. They found that a surprisingly high number of 42 examples (76%) were quotations, which their system did not actually try to identify. Most of these 42 quotations were by persons whose name was not mentioned at all in the article (e.g. the officer / their neighbor). The remaining ones were by persons that are not part of their known persons. For the remaining 13 cases, i.e., those that do fall inside their mandate and that they do try to identify, seven were correct while 6 had not been found, corresponding to a Recall of 54%.

de La Clergerie et al. (2009) developed a system called SAPIENS, which extracts quotations from news wires, associated with their author and context. SAPIENS was applied on a corpus of news wires from the Agence France-Presse. The AFP produces every day 6,000 news in six languages. The average size of a news item is 250 words.

Their system is composed of the following modules: named entities extraction, verbatim extraction, deep parsing, anaphora resolution and quotation. They claim that the information made available by SAPIENS is richer and more accurate than other systems such as Google InQuotes, mainly because of the use of a deep parser. Their named

entity detection module has been developed in SXPipe's dag2dag framework (Sagot and Boullier, 2008). Disambiguation heuristics have been activated, so that the amount of ambiguities added by this named entities module is as low as possible.

3 QUOTATION STYLES

3.1 Quotation Styles in General Text and Scientific Papers Written in English

There are various styles of quotations. These styles depend mainly on the length of the quotation. For instance:

(1) Quoting of relatively short paragraphs (less than 35 words or less than 50 words or less than five lines) - the quotation is enclosed by a pair of quotation marks (e.g., “, ”, ', ', «, »), and the parenthetical citation should be after them and before the period. Another method (but less popular) is to write the quoted words using the italic type.

(2) Quoting of larger paragraphs - the quotation need to start on a new line. The entire quote should be indented from the left side of the regular text. The citation should appear after the period. Quotation marks and/or italic type are not needed.

(3) Quoting of a quotation that contains quotation - alternate the use of double and single quotation marks for the various quotations.

3.2 Quotation Styles in Hebrew-Aramaic Documents

Structural lack of rabbinic texts written in Hebrew-Aramaic is a complex challenge. Quote identification is not trivial, because it is not necessarily marked in the rabbinic text with quotes, italic writing or indentation as customary in scientific texts. Therefore, an incorrect quote (a short one, usually) might be retrieved even when the author did not mean to quote.

As mentioned before, Hebrew quotations can be presented with different types of prefixes included in the name of a citation.

3.3 Examples for Quotations in Hebrew-Aramaic

In this sub-section, we detail five examples. Two of them are positive sentences (or clauses) that each includes one Biblical quotation and three negative sentences (or clauses) that each does not include any Biblical quotation.

Examples of Positive Sentences: In the first example, the quotation is enclosed by a pair of quotation marks, while the second example does not include a pair of quotation marks.

(1) נח נקרא נח משום שעליו נאמר "זה ינחמנו ממעשינו ומעצבון ידינו"

Noah was called Noah because on him was said "The one who will bring relief from our work and from the toil of our hands" (Genesis 5, 29).

(2) ... נפקא ליה מומקלל אביו ואמו מות יומת ...

... from the verse: one who curros his father or his mother shall surely be put for death … (Sanhedrin, chapter ten, 85b).

Examples of Negative Sentences: The first example contains a quotation that is enclosed by a pair of quotation marks; however this quotation is not Biblical. The second example includes a quotation which is not Biblical and is not enclosed by a pair of quotation marks. The third example does not contain any quotation.

(1) שמעון בנו אומר כל ימי גדלתי בין החכמים ולא מצאתי לגוף טוב משתיקה ...

Shimon his son said "all my days I have been raised among the Sages and I found nothing better than silence …" (Pirke Avot 1, 17).

(2) משה אמר שעודד אמר שהוא לא יבוא

Moshe said that Oded said that he will not come.

(3) רונן לא מתכוון לבוא מחר

Ronen does not intend to come tomorrow.

4 FEATURE SETS FOR QUOTATION IDENTIFICATION

Nineteen different features have been defined for the identification of quotations included in rabbinic documents. These features were divided into seven different feature sets: matches, best matches, sums of weights, weighted averages, weighted medians, common words, and quotation indicators. All feature sets are domain-independent and language-independent.

Unfortunately, no POS-tagger for Hebrew-Aramaic texts was available to us. There are two online available Hebrew taggers. However, these taggers are suitable to deal with Modern Hebrew texts, which are quite different from rabbinic Hebrew-Aramaic texts that we are dealing with.

The feature sets together with their features are presented below.

"Matches" (M) Features:

1. **MC-A: Count of All Matches** (partial and complete) between the tested quotation and various Biblical sources.
2. **MC-C: Count of Complete Matches** between the tested quotation and various Biblical sources.
3. **MC-L:** Count of Lacking Matches between the tested quotation and various Biblical sources.

"Best Matches" (BM) Features:

4. **BM-A:** The weight of the Best Match alone.
5. **BM-M:** The weight of the Best Match Multiply by the Matches Count (MC-A).

"Sum of Weights" (SW) feature:

6. **SW-A:** Sum of All matches' Weights.

"Weighted Averages" (WA) Features:

7. **WA-A:** Weighted Sum of All matches.
8. **WA-F7:** Weighted Sum of First 7 matches.
9. **WA-F3:** Weighted Sum of First 3 matches.
10. **WA-S3:** Weighted Sum of Second 3 matches.
11. **WA-T3:** Weighted Sum of Third 3 matches.

"Medians" (WM) Features:

12. **WM-A:** The Median value of All matches.
13. **WM-F7:** The Median value of First 7 matches.
14. **WM-F3:** The Median value of First 3 matches.
15. **WM-S3:** The Median value of Second 3 matches.
16. **WM-T3:** The Median value of Third 3 matches.

"Common Words" (CW) Features:

17. **CWC-CP:** Count of Common Words between the tested quotation and the Cited Part of the Biblical source divided by the number of words in the Cited Part of the Biblical source.
18. **CWC-AS:** Common Words Count – Count of Common Words between the tested quotation and the whole Biblical source divided by the number of words in the Cited Part of the Biblical source.

Quotation Indicator (QI) Feature:

19. **QI-QM:** Quotation marks. This feature examines whether the tested sentence includes a pair of quotation marks. If the answer is yes than the result of this feature is that this sentence includes a quotation, otherwise - it is does not include a quotation.

5 THE MODEL

The main steps of the model are presented below:

1. Building a corpus containing 2,000 sentences: half of them containing a single quote (positive sentences) and the other half containing no quotes (negative sentences)
2. Cleaning the sentences. We delete various types of additions in brackets and redundant spaces that are not part of the original text.
3. Calculating the 19 features for each sentence.
4. For each tested combination of feature sets (each feature alone, a pseudo backward hill climbing method and a pseudo forward hill climbing method) apply each chosen supervised ML method using features chosen according to InfoGain's[1] estimates using a 10-fold cross-validation.

6 EXPERIMENTS

To accomplish the task of determining whether there is or there is not a quote in a particular sentence a corpus was built, containing 2,000 sentences. To prevent biased learning, these sentences were selected in such a way that they are divided into two equal-sized sets. 1,000 sentences are positive examples (i.e., each one of them includes one quotation) and 1,000 sentences are negative examples (i.e., each one of them does not include any quotation at all). To the best of our knowledge, there are no false examples in the training data.

The accuracy that was measured in all experiments is the fraction of the number of sentences that the quotations in them were correctly identified to the total number of positive sentences.

Four well-known supervised ML methods suitable for text processing have been selected for the application of the last stage in our model: Logistic Regression (LR) (Hosmer and Lemeshow, 2000), Multi-Layer Perceptron (MLP) (also known as an Artificial Neural Network) (Haykin, 1998), SMO (Platt, 1999) (a variant of the SVM method (Cortes and Vapnik, 1995; Vapnik, 1995)) and J48 (an improved variant of the C4.5 decision tree induction (Quinlan, 1993)). These methods have been applied using Weka (Witten and Frank, 2009). To test the accuracy of the model, a 10 times 10-fold cross-validation was performed.

[1] We choose InfoGain since it is a very successful feature selection metric (Yang and Pederson, 1997; Forman, 2003).

Table 1: Accuracy results for each feature using the ML method.

#	Feature Set	LR	MLP	SVM	J48
1	**MC-A**	81.6	81.65	70.8	82
2	**MC-C**	80.6	81.05	75.2	80.9
3	**MC-L**	78.05	78.6	70.65	78.7
4	**BM-A**	81.9	81.9	81.9	81.9
5	**BM-M**	81.55	81.6	69.9	82.15
6	**WS-A**	81.5	81.55	69.25	82.05
7	**WA-A**	81.9	81.4	79.75	81.6
8	**WA-F7**	81.9	81.4	75.4	81.85
9	**WA-F3**	81.9	81.45	78.45	82.15
10	**WA-S3**	81.6	81.6	81.6	81.6
11	**WA-T3**	76.55	76.55	70.8	76.55
12	**WM-A**	74	72.15	74	74
13	**WM-F7**	62.8	61.85	62.8	62.8
14	**WM-F3**	62.8	61.8	62.8	62.8
15	**WM-S3**	81.6	81.6	81.6	81.6
16	**WM-T3**	76.55	76.55	76.55	76.55
17	**CWC-CP**	65.9	65.95	67.05	68.9
18	**CWC-AS**	66.4	65.75	65.1	66.6
19	**QI-QM**	83.15	83.15	83.15	83.15

The four right columns of Table 1 show the accuracy results for each combination of one feature and one ML method. The best result has been achieved by QI-QM (#19). This feature examines whether the tested sentence includes a pair of quotation marks. Testing this feature on the examined dataset yields an accuracy of 83.15%.

Since many of the quotations are enclosed by a pair of quotation marks, it was rather logical to choose the QI-QM feature as our baseline identifier. Among the other features, ten features (MC-A, MC-C, BM-A, BM-M, WS-A, WA-A, WA-F7, WA-F3, WA-S3, and WM-S3) achieve accuracy rates between 80.9% to 82.15%.

Table 2: Accuracy results for combinations of features using a pseudo backward hill climbing method.

Removed Features	#	LR	MLP	SMO	J48
-	19	90.55	90.75	90.4	90.85
WM-F3	18	90.55	90.75	90.4	90.85
WM-F7	17	90.55	90.6	90.4	90.75
CWC-CP	16	90.6	90.8	90.4	90.7
CWC-AS	15	90.55	90.55	90.4	90.4
WM-T3	14	90.45	90.5	90.4	90.4
WM-A	13	90.5	90.45	90.4	90.4
WA-T3	12	90.45	90.6	90.25	90.4
WM-S3	11	90.4	90.65	90.5	90.4
QI-QM	10	82.15	81.6	81.6	81.95

Table 2 presents the accuracy results for combinations of features using a pseudo backward hill climbing method. At the beginning we tried all 19 features. Then, in each step, we omitted the feature with the lowest weight according to InfoGain. Table 2 shows the results of each combination of features with at least 10 features using this method for each ML method.

The best accuracy result (90.85%) was achieved by J48, based on 18 (also 19 features, but 18 is, of course, better) features using the pseudo backward hill climbing method. The second best result (90.8%) was achieved by MLP, based on 16 features.

Table 3 presents the accuracy results for combinations of features using a pseudo forward hill climbing method. We started with the best feature, the QI-QM feature with accuracy result of 83.15%. In each step, we added the feature with the highest weight according to InfoGain. Table 3 presents the results of each combination of features with at most 4 features using this method for each ML method.

Table 3: Accuracy results for combinations of features using a pseudo forward hill climbing method.

#	Chosen Features	LR	MLP	SMO	J48
2	WA-A,QI-QM	89.05	89.45	89.25	90.2
	WA-F7,QI-QM	89.85	90.15	89.85	90.15
	WA-F3,QI-QM	90.15	90.55	90.5	90.4
	BM-A,QI-QM	89.7	90.05	90.05	90.05
	BM-M,QI-QM	90.35	90.35	87.45	90.45
3	MC-A, BM-M,QI-QM	90.05	90.45	88.85	90.45
	BM-M, WA-S3,QI-QM	90.3	90.05	90.25	90.45
	BM-M,WM-S3,QI-QM	90.45	89.95	89.1	90.45
	MC-A, WA-F3,QI-QM	90.45	90.4	90.5	90.4
	WA-F3, WA-S3,QI-QM	90.15	90.55	90.25	90.4
	WA-F3,WM-S3,QI-QM	90.15	**90.6**	90.55	90.4
	WS-A, BM-M,QI-QM	90.5	90.4	88.8	90.4
4	WS-A, BM-M,WM-S3,QI-QM	90.5	90	89.1	90.45
	WS-A, WA-S3,WM-S3,QI-QM	90.5	90.2	90.2	90.45
	WS-A, WA-A,WM-S3,QI-QM	90.4	90.1	89.1	90.4
	MC-A, BM-M, WA-S3,QI-QM	90.5	90	90.25	90.5
	MC-A, BM-M,WM-S3,QI-QM	90.5	90	89.1	90.45

The best accuracy result (90.6%) in table 3 was achieved by MLP, based on 3 features.

The pseudo backward hill climbing method achieved a slightly better result (90.85%) than the best result (90.6%) achieved by the pseudo forward

hill climbing method. Therefore, we decided to focus on the pseudo backward hill climbing method regarding parameter tuning. We tried a widespread variety of experiments with all the four ML method. Due to space limitations, we will detail the best result, which was achieved by J48.

We decide to limit the number of leaves in each sub-decision tree to 18. By that, we prevent over fitting. In addition, we prevent any pruning. Features # 11 & 12 that their weights were zero according to InfoGain were omitted. By that, we achieved 91.2%. After the omitting of the third feature, the accuracy rate fall to 91.05%

To sum up, the best accuracy result (91.2%) was achieved by J48, based on 17 features using the pseudo backward hill climbing method. It represents an improvement of 8.05% to the baseline result (83.15%).

7 SUMMARY AND FUTURE WORK

We present an automatic process that identifies quotations included in rabbinic documents written in Hebrew-Aramaic. The identification was based on a combination of at most nineteen features that belong to seven feature sets: matches, best matches, sums of weights, weighted averages, weighted medians, common words, and quotation indicators.

Our research is unique in addressing the much more difficult problem of identification of quotations included in rabbinic literature. Furthermore, we define and apply features that some of them have not been used in previous researches.

Experiments on various combinations of these features were performed using four common ML methods. A combination of 17 features using J48 (an improved version of C4.5) achieves an accuracy of 91.2%, which is an improvement of about 8 % compared to a baseline result.

Other Semitic languages, such as Arabic also have similar complex morphology. For example, prefixes and terminal letters might be included in the citations. Research that emphasizes morphological features might be fruitful not only to Hebrew-Aramaic rabbinic texts but also to other kinds of texts from other Semitic languages.

Future research proposals are: (1) Identify more complex cases, e.g., nested quotations and quotations attributed to unspecified authors; (2) Implement morphological features that are appropriate for various Semitic languages; and (3) Disambiguate ambiguous quotations.

REFERENCES

Choueka, Y., Conley E. S., Dagan. I., 2000. A Comprehensive Bilingual Word Alignment System: Application to Disparate Languages - Hebrew, English, in *Veronis J. (Ed.), Parallel Text Processing*, Kluwer Academic Publishers, 69-96.

de La Clergerie É., Sagot, B., Stern, R., Denis P., Recourcé G., Mignot. V., 2009. Extracting and Visualizing Quotations from News Wires, in Proc. of L&TC 2009, Pozna´n, Poland.

Cortes, C., Vapnik. V., 1995. *Support-Vector Networks*. Machine Learning 20, 273-297.

Forman, G., 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. of Machine Learning Research* 3 1289-1305

Gabrilovich, E., Markovitch, S., 2004. Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5. In *Proc. of the 21 International Conference on Machine Learning*, 321-328, Morgan Kaufmann

Haykin, S., 1998. *Neural Networks: A Comprehensive Foundation*, 2nd edition, Prentice Hall.

Hosmer D. W., Lemeshow. S., 2000. *Applied Logistic Regression*. 2nd ed. New York; Chichester, Wiley.

Liang, J., Dhillon, N., Koperski. K., 2010. A Large Scale System for Annotating and Querying Quotations in News Feeds, *Semantic Search 2010 Workshop ,the 19th international conference on World wide web (WWW2010)*, Raleigh, North Carolina, USA.

Miller, G., A., 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity of Information. *Psychological Science*, 63, 81-97.

Platt, J., C., 1999. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, MIT Press, Cambridge, Massachusetts, chapter 12, 185-208.

Pouliquen, B., Steinberger, R., Best, C., 2007. Automatic Detection of Quotations in Multilingual News. In *Proc. of Recent Advances in Natural Language Processing (RANLP-2007)*, 25-32.

Quinlan. R., J., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos.

Sagot, B., Boullier, P., 2008. SXPipe 2: Architecture pour le Traitement Présyntaxique de Corpus Bruts. Traitement Automatique des Langues (T.A.L.), 49(2):155-188.

Yang, Y., Pedersen, J., P., A., 1997. Comparative Study on Feature Selection in Text Categorization. In *Proc. of the Fourteenth International Conference on Machine Learning (ICML'97)*, 412-420.

Vapnik. V., N., 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag, NY, USA.

Witten, I., H., Frank, E., 2009. *Learning Software in Java*. http://www.cs.waikato.ac.nz/~ml/weka.