Modeling the Hebrew Bible : Potential of Topic Modeling Techniques for Semantic Annotation and Historical Analysis

Mathias Coeckelbergs,
Seth van Hooland

Université Libre de Bruxelles
Département des Sciences de l'information et de la communication
Avenue F. D. Roosevelt, 50 CP 123
B-1050 Bruxelles, Belgique
mcoeckel@ulb.ac.be - svhoolan@ulb.ac.be

Abstract

Providing useful and efficient semantic annotations is a major challenge for knowledge design of any body of text, especially historical documents. In this article, we propose Topic Modeling as an important first step to gather semantic information beyond the lexicon which can be added as annotations in the SHEBANQ. By laying out a case study, we discuss both noise and structure found in comparing topics extracted within different distributions, and show the value of such approach, which we label a topic hierarchy. We also show a first result in applying such approach to study diachronic variety in the Bible, and show how this overall Topic Modeling approach can result in more query options for users of the database.

Keywords
Historical Ontology, Natural Language Processing, Semantic Annotation, SHEBANQ, Topic Modeling

1 Introduction

1.1 The Relationship between Hebrew Studies and Natural Language Processing

The Hebrew Bible has been one of the central building blocks of scholarship across the history of the humanities, adding layers upon layers of interpretations and comments across languages and time [1]. Due to its importance within the humanities and the sheer volume of additional information and interpretations revolving around the Hebrew Bible, scholars of ancient literature and linguistics have been early adaptors of computational methods to develop new research questions. The usage of computational methods has first and foremost resulted in the digitisation of the most relevant texts, including the Hebrew Bible itself and the Dead Sea Scrolls (DSS) for example, but also in the application of Natural Language Processing (NLP) on mainly Greek and Latin sources. The Perseus project[1], which focuses on both primary and secondary source materials related to the Greco-Roman world, is

[1] http://www.perseus.tufts.edu

probably one of the most well-known projects. Other projects focus on the annotation of texts with linguistic tags allowing in-depth queries into the surface structure of the text, such as for example the Bibleworks 10[2] software which contains lexical, morphological and accentual tags. Syntactic tags are not available here, but can be found for example in Accordance[3] and Logos[4], as well as in SHEBANQ[5]. This recent database, developed by the Eep Talstra Centre for Bible and Computer, contains the corpus and the metadata which are the focus of the work presented in this paper. These projects have since their inception seen an increasing uptake in the scholarly community, albeit mainly for *close reading* practices. This term refers to the traditional humanistic method of reading and studying in detail a specific part of a corpus, which generally stems from a well delimited academic canon. This article is to be placed in the general project of moving towards *distant reading* methods for the study of the Hebrew Bible, as developed by Moretti [2]. Projects such as Perseus and SHEBANQ facilitate detailed reading and interpretation of texts due to the ease of access to the documents and the hyperlinks to grammars, dictionaries and concordances. However, recent developments from the Semantic Web and the NLP communities hold the promise of going well beyond a mere presentation of texts and hyperlinks for human consumption.

1.2 Research corpus: SHEBANQ

The System for HEBrew text: ANnotations for Queries and Mark-up (SHEBANQ)[6] constitutes our core corpus [3]. This is an online platform allowing a systematic study of the Hebrew Bible, centered around the idea of user annotations, enriching the text with a multitude of additional information, most notably queries. Until now, interest has mainly focused on expanding the syntactic descriptive granularity of the text. The long-term goal of the project is to create an ever-expanding platform including a critical apparatus of metadata annotations which need not be limited to syntactic data. The logical next step is to explore possibilities of adding semantic information surpassing the lexical level. For the given corpus, this is a daunting task due to the enormous variation of texts extant in the Hebrew Bible, which hamper simple application of current text mining techniques on the semantic level such as Named Entity Recognition (NER) and Topic Modeling (TM).

As already stated, the main databases apart from the SHEBANQ are the Logos Bible Software, Accordance and Bibleworks. Together they constitute a useful digital source of the main Hebrew texts, also non-biblical, concerning morphology, lexicon and syntax. These texts can easily be exported to other document extensions such as .doc or .txt, but a general unicode version of the texts can be found on the site of the German Bible Society[7]. Concerning the semantics of the Biblical text, important work is being done by Reinier De Blois in the Semantic Dictionary of Biblical Hebrew[8]. This project does not focus on the historical development of word meaning, a task taken up by the Historical Dictionary Project at the University of Jerusalem[9]. Scholarly literature on these issues are gathered by the Semantics of Ancient Hebrew Database[10].

An important improvement of the SHEBANQ in comparison to other Hebrew semantic information is its novel way of representing data. For this, it uses the Lin-

[2] http://www.bibleworks.com
[3] http://www.accordancebible.com
[4] https://www.logos.com
[5] https://shebanq.ancient-data.org
[6] Ibid.
[7] https://www.dbg.de/
[8] http://www.sdbh.org/
[9] http://hebrew-academy.huji.ac.il/English/HistoricalDictionaryProject/
[10] http://www.sahd.div.ed.ac.uk/

guistic Annotation Framework [4]. Its main asset for our purposes is the possibility of stand-off markup, whereby the primary text is left untouched and annotations are added as feature structures. The whole file then is structured as a graph, consisting of nodes and edges. Nodes are assigned to every individual constituent of the file, both in the primary text as well as in the annotations. Edges are used to connect nodes into an ever larger picture, linking relationships between constituents and entities.

2 Integration of the Hebrew Bible within the Semantic Web community

2.1 General Project and the Role of Topic Modeling

The Topic Modeling approach goes both beyond the level of surface forms and beyond problems of spelling differences for example, to a more abstract level of semantic entities that cannot be readily found by straightforward textual search. Building salient models for a given corpus is not an easy task, as hallmarked for example by [5] [6] [7]. A primary task will consist in finding adequate models for each of the Biblical books, which diverge enormously regarding length, content and style. Hence, we will have to study the importance of the amount of topics per book, manual and automatic assignments of topics, probability of co-occurrence of terms in topics of varying length, and other questions regarding the basic architecture of our model.

Once we have gathered the necessary information on the topic architecture, we can proceed to annotating these data into the SHEBANQ. Several possible ways exist to proceed in this task. One viable way consists of annotating the confidence factor with which a word or sentence belongs to a certain topic as metadata, and work out how topic assignment on a higher level (paragraph or an entire book) can proceed from this information. Another possibility of annotating is to focus on the other words that receive high probability of co-occurring with a given word in a certain topic. Given a word, related words can be annotated, so that discourse structures can be evaluated on the significance with which they represent a topic.

The purpose of annotating topics shares an overlap with that of a concordance, namely to have a concrete idea of the contexts in which a certain word appears. The first digital version was famously created by Father Roberto Busa S.J., with his monumental work on the *Corpus Thomisticum*[11]. This is a searchable database of all words written by Thomas Aquinas and related authors, constituting a resource of about 11 million words. Classical concordances, either on paper or digital, present context and use of words by listing an (exhaustive) list of sentences representing the semantic range of a concrete word. Our annotated topics will not place the word in reference to concrete sentences, but to a probability with which they co-occur with other words within several topic distributions.

This addition of TM to the database should result in the improvement of tasks readily available, such as for example query expansion. With this step, we wish to contribute to the analysis of TM as a method for data discovery, which has been both received positive and negative critiques. Positively, it allows for new ways of comparing and discussing texts, whereas negatively these possibilities may be overestimated to form misleading conclusions. Problems lie in the interpretation of clustered words together as constituting a coherent topic, whereas this is not necessarily always so clear-cut. With our topic hierarchy method, described below, we wish to address this issue in a novel way. Continuing with this approach, we want to outline as well how we can use this extracted information to automatically

[11] `http://www.corpusthomisticum.org/`

create authority data which can be used to link the data in our database to other, related information. This can be contained in a variety of resources, including most notably research articles, online fora, blogs and social media. The JSTOR Labs API [12] allows us access to a fuzzy text matching algorithm for linking secondary literature citations of Shakespeare to the original text. A similar approach is also possible for the Hebrew Bible, which can be linked to our topic hierarchy. Our main purpose will be to allow users of the Bible the highest level of querying freedom as possible, in order to contribute to a digital critical edition of the Bible which uses current text mining tools to their fullest potential. In the end, the TM architecture described in the next chapter constitutes a firm basis as pre-training before applying NER to the corpus, as done recently in for example [8] This is an important asset for a contribution to this recent field of inquiry.

2.2 Proof of Concept

In this subsection we describe our first attempts of building a topic hierarchy, by extracting different topic distributions from the Hebrew Bible and comparing their inherent relationship and usefulness. As a training set we used the book of Genesis, a relatively large book which incorporates both narrative and poetic parts, to test the LDA Algorithm on the other books of the Hebrew Bible. The study of model making on the basis of ancient Hebrew literature, and more specifically the importance of adapting training and test data to each other are in need of further elaboration still, and our results may be skewed because of our particular choice. For the execution of this algorithm, we rely on the Mallet software[13] for its ease of adaptation of feature values. Of course, further in the future other software and algorithms will have to be subject of investigation as well. Tests comprised the creation of several topic models of varying amount of topics (a so-called distribution) to be found in the text. In accordance with our expectations, the fewer topics that needed to be found in the text, the less coherent the topics seem to be. As a topic model is nothing more than a group of words which are probabilistically clustered together, human interpretation of these clusters is a problem in itself, an issue which is out of scope of the present article.

Looking at our first results comparing different topic distributions (going from 30 to 90 topics with increments of 10), we notice both chaos and interesting structure[14]. The more words two topics across distributions have in common, the darker the line connecting them is coloured. In figure 1, we have filtered out lines connecting only a few words in order to preserve clarity. We see that some topical structures are apparent in all distributions, sharing a remarkable amount of words. This is for example the case for topics concerning kingship. Common words for this topic

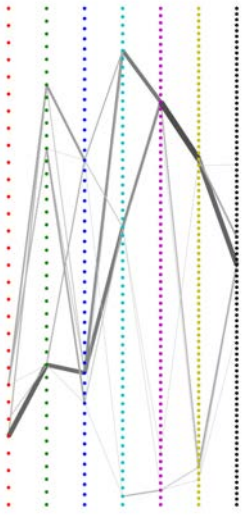

Figure 1: Most important similarities of topic distributions ranging from 30 to 90 topics with increments of 10

[12] `https://labs.jstor.org/apis/docs/`
[13] `http://mallet.cs.umass.edu/`
[14] We would like to thank Dirk Roorda for his help with visualising topic hierarchies

include Hebrew words for *king (mlk)*, conjugated forms of the verb *to rule (mlk)* and *throne (ks')*. Small distributions also include different named lands, such as Edom and Israel, which appear to be filtered out in higher ones. This can be explained because low level distributions still contain clustered words with a relatively large distance to the cluster centre.

Looking at the noise we found in our dataset, further investigation of primary results will allow us to finetune the model and decrease the amount of noise. One of these points of inquiry is to find a balance between treating every occurrence of words (in all inflections, with conjunctions, etc.) as separate words, or dealing only with lexemes. The latter shows finer clusters at some points, for example clustering more finegrained topics, containing more specific lexemes, concerning kingship, already at a lower distribution. On the other hand, finding meaningful clusters is more difficult on the lexeme level, because words occurring in similar contexts, derived from the same lexeme, can no longer be clustered together, which would provide topics easily recognisable for humans. Next to considering the relation between words and lexemes, we must also focus on removing other words with low semantic value, such as conjunctions, adverbs and the like, because they result in less meaningful clusters when taken up. This is more difficult for Hebrew than for English, because for example the conjunction *waw* is added to the beginning of the word, never appearing separately. We cannot simply break this loose from the word it is attached to, because in Hebrew, this conjunction is also part of the pragmatic value of a sentence, making it hard to treat both verbs, with and without conjunction, as equal. A final point we have to further investigate is how words from coarser topics in low level distributions are reclustered in higher level distributions. For example, we can see that a low level topic related semantically to both fighting and dying is in higher distributions split up in a topic dealing mainly with words revolving around fighting, and another with only words in the semantic field of dying.

2.3 Collecting Vocabulary for the Preparation of an Ontology

As stated in our introduction, we believe that our Topic Modeling approach can assist in developing new insights into the diachrony of the Hebrew literature, which adds important information for an historical ontology of Hebrew literature. Early results show that typical words for a certain time period in Hebrew tend to be clustered together if they are low in semantic meaning, such as conjunction words. Results to what this can mean for shifts in concepts such as kingship are still ongoing, and it is as yet too early to draw primary conclusions. Expected improvement is situated mainly on differences between extracted topics dealing with a similar content from all books of the Bible independently, in comparison to topics extracted from the entire Bible.

The value of our approach for this enquiry can be seen in for example the evolution of poetic writing, the different ways of speaking about enemies before and after exile, and the reception of Israelite history in the books of kings in comparison to the books of Chronicles. General scholarship accepts the view that the latter is a rewriting in later times of the former. This is a unique asset for studying the diachronic development of language use, which can be seen in change in vocabulary and orthography, but also in the evolution of the semantic field used to speak about key issues such as kingship, victory and rules. However, to date no clear description of this intricate development has been shown using modern tools.

3 Conclusions and Future Work

In this article, we hope to have contributed to a new way of investigating Hebrew literature through the use of Topic Modeling. The basic methodology and architecture of our approach was described and we pointed out the possibilities of how this approach can help to detect historical relations within the text. Future work will be situated on two fronts. On the one hand, we will use previous work done at JSTOR labs linking works of Shakespeare and their scholarly work, to allow a similar approach to the Hebrew Bible, and study how this knowledge can be used in conjunction with our topic hierarchy to allow the user more query options. On the other hand, we will try to improve topic extraction from the Bible, discussing the relationship between concrete words versus lexemes, the similarity between similar topics from different distributions, and several smaller levels of scope to perform Topic Modeling on, such as individual Bible books, chapters, or other meaningful units.

References

[1] Bod, R.: A New History of the Humanities. The Search for Principles and Patterns from Antiquity to the Present. Oxford University Press, 2013.

[2] Moretti, F.: Graphs, Trees, Maps. Verso, 2005.

[3] Roorda, D.: The Hebrew Bible as Data: Laboratory, Sharing, Experiences. ArXiv:1501.01866.

[4] Eckart, K.: A Standardized General Framework for Encoding and Exchange of Corpus Annotations: The Linguistic Annotation Framework, LAF. In: Proceedings of KONVENS, GAI, pp.506-515 (2012).

[5] Efron, M., Organisciak, P., Fenlon, K.: Building Topic Models in a Federated Digital Library through Selective Document Exclusion. In: Proceedings of the American Society for Information Science and Technology vol. 48, pp. 1-10 (2011).

[6] Brauer, R., Friedlund, M.: Historicizing Topic Models. A Distant Reading of Topic Modeling Texts within Historical Studies. In: International Conference on Cultural Research in the context of ?Digital Humanities?, pp. 152-163 (2013).

[7] Mimno, D.: Computational Historiography: Data Mining in a Century of Classics Journals. In: Journal of Computing and Cultural Heritage, vol. 5, pp. 3-22 (2012).

[8] De Wilde, M., Hengchen, S.: Semantic Enrichment of a Multilingual Archive with Linked Open Data, http://dhbenelux.org/wp-content/uploads/2015/04/06.pdf .