Modeling the Hebrew Bible : Potential of Topic Modeling Techniques for semantic annotation of the SHEBANQ

Mathias Coeckelbergs
Seth van Hooland

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

1 SHEBANQ and its potential for annotating scientific texts

In recent years, two fairly separated research fields have made important improvements and grew closer together. On the one hand, several researchers in computer science and information science have turned their attention to ancient material. Among these, mainly Greek and Latin sources have undergone natural language treatment going beyond mere digitization of the extant textual resources. Probably the best know example of these efforts is the Perseus project. On the other hand, scholars in the study of ancient literatures and linguistics have acknowledged the importance of computational methods to propose new questions and answers in their research field. This insight has resulted in the digitization of the most relevant cultural texts, including for example the Hebrew Bible and the Dead Sea Scrolls. For these two important artefacts of the Hebrew language, we already have several computer programs available containing databases of text richly completed with linguistic tags. The Bibleworks 10 software contains lexical, morphological and accentual tags. Syntactic tags are not available here, but can be found in Accordance (the database organized by R. Holmstedt) and Logos (organized by F. Andersen and D. Forbes), as well as in SHEBANQ, the recent database organized by the Eep Talstra Centre for Bible and Computer. These tools have since their inception seen an increasing use in the scholarly community, albeit mainly for close reading it [1], due to the ease of access to the documents and the hyperlinks to relevant works such as grammars, dictionaries and concordances. We will build upon the data already available in these sources, and, where needed, we will provide additional parsing for optimal descriptive granularity, focusing mainly on the improvement of the latter database, which has been specifically designed to allow further research with natural language processing tools.

The SHEBANQ (System for HEBrew text: ANnotations for Queries and Markup) [2]database will be the main focus of our research. This is an online system allowing systematic study of the Hebrew Bible, which is centered around the idea of allowing users to annotate 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. Goal of the enterprise is to create an ever-expanding database including a critical apparatus in metadata annotation 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 mining techniques on the semantic level such as NER, TM, etc. Extant texts include a wide variety of genres, including narrative, legal and poetic, different lengths of text ranging from merely 21 verses (Obadja) to 2461 (Psalms), diachronic variety of texts comprising several centuries on morphological, syntactic, semantic and orthographical level. Since the specific research question guiding our project is to critically analyze extant methods for extracting semantics of large volumes of non-structured text, we will pinpoint our approach within the history of data models.
Computational methods for the DSS are only recent and rare, but for the Hebrew Bible first interests lay already in the 1970s. However, this computational interest has remained limited to the use of quantitative arguments, such as for example the (non)occurrence of particular lexemes in classifying texts, leaving more elaborate methods unapplied to a Hebrew corpus. Moreover, many of these arguments are methodologically flawed, remaining on the level of observing frequencies without taking into account their significance or failing to acknowledge the limited set of features under consideration, raising questions concerning the conclusions of these results, indicated by quantitative arguments for opposite views. Text mining methods can contribute to the desideratum of computationally approaching the semantics of the Hebrew Bible, such as named entity recognition or word sense disambiguation. In the SHEBANQ, several TM tools have already been implemented to broaden the applicability for users.

2 Topic Modeling and the Language Annotation Framework

A main problem of historical texts is the challenge of language evolution, resulting in differences in vocabulary, language structure, and specifically for Hebrew, a difference in use of matres lectionis (reading cues guiding the interpretation of vowels, which are not explicitly written in Hebrew). Topic modeling goes beyond the level of surface forms, leaving behind problems of spelling differences for example, to a more abstract level of semantic entities that cannot be readily found by straightforward textual search. Our first task of topic modeling 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. It will be our goal to experiment with different possibilities and conclude which technique is most appropriate. 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 would focus on the other words that receive high probability of co-occurring with a given word in a certain topic. Given a word, we can annotate these related words, so that discourse structures can be evaluated on the significance with which they represent a topic. This addition of topic modeling 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 topic modeling as a method for data discovery, which has been proposed in several recent articles as a valuable road ahead for research in the applicability of this technique. Up until now, its use was generally limited to document classification, relinquishing its capabilities for discovering new information and links between data. Continuing in this same strand, we want to focus in our research project to use this information to create a ontology which can be used to link the data in our database to other, related information (research articles, online fora, blogs, social media,).

3 Proposed Semantic Annotation of Topics

We have worked out a first tentative test case that can point out the usefulness of the research proposal sketched above. As a test set we took the book of Genesis, an often selected corpus for first trials, with models trained on the text of the entire Bible. This selection of training data may skew results, although the test set represents a similar distribution of texts as the training set, incorporating both narrative as poetic parts. 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. Using the Mallet software, we performed topic modeling on our corpus via the Latent Dirichlet Allocation algorithm. Our tests comprised the creation of several topic models of varying amount of topics to be found in the text. In accordance with our expectations, the less topics that need to be found in the text, the less coherent the topics seem to be. This is due to the ever increasing amount of words in the same cluster. 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, which we will leave untouched here. However, it is apparent that in comparing the co-occurrence of two not seldom attested words in the same topic among several topic distributions we have made on top of the text, some words clearly appear to have higher probability of occurring together than others. Although this conclusion is trivial, we can use its premise to build a model to quantify the strength with which words are connected to each other. We will discard concrete mathematical properties for now, since this test case is at this point only tentative. In the next two paragraphs we will work out the architecture of our idea for semantic annotation based on topic modeling, and discuss its possible extensions for insight in the diachronic

development of the text.

3.1 Architecture

The purpose of annotating topics shares an overlap with that of an accordance, namely to have a concrete idea of the contexts is which a certain word appears. Classical accordances do this 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 process should make it possible to link words more easily to other relevant words, given the topic of the initial one. Within the context of our corpus this will primarily result in more efficient query expansion and query results. Furthermore, this might also facilitate the still open question of linking relevant scholarly work and other secondary sources to words in the primary text.
Let us start with a concrete example to illustrate the possibilities of this concept. Using the LDA algorithm to find twenty topics in the book of Genesis, we have found the word ishah (woman) to belong to a topic which contains other words such as tachat (below), wmat (and he died), adam (man), nachash (snake), Yaakov (Jacob). Allowing the algorithm to create more topics, we see the word ishah connected to still fewer words. Having 35 topics, we see for example that it is still connected to tachat, adam and nachash, no longer to Yaakov and wmat. Hence, the former group has a stronger likelihood of belonging to the same topic that the latter one. After further increasing the granularity of topics, we find that among the five original words, adam is linked most strongly to ishah, probably due to the creation epic of Genesis II and the abundant talk of bonding and marriage between man and woman throughout the book.
For our concrete work with the SHEBANQ, we propose to annotate the identity card of words appearing in topic models in distinct topic distributions. Via a still further to specify mathematical relation, the confidence factor of a word belonging to a certain topic in a certain topic distribution will be visible after words appear as results for a query. This in turn can be used to rank several other verses in our corpus according to relevance for the topic of the search query. We hope that this procedure can help to address new research question concerning the corpus, and that it will prove to better visualize its structure.

3.2 Possibilities for enhancing the building of a historical ontology

Our expectation is that the previous approach can be elaborated further to help improve a knowledge base to build a historical ontology, in the process of which our knowledge of the diachronic development of the Hebrew texts of the Bible, as well as evolutions in vocabulary and orthography can be improved. We will discuss briefly these three possibilities of gaining insight development of the texts which can be accommodated in an ontology. Firstly, topic modeling can give us insight into the shift in talking about certain key ideas throughout the development of the Bible. This can be considered in many instances, 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.

Secondly, delving deeper into the problem of developing vocabulary, we can consider hapax legomena, words which are only once attested in our corpus. These words form a high point of interest among hebraists, because the meaning of these words is contested in general. Suggestions for their meaning are usually proposed by considering cognate words in related languages (most notably Aramaic and Ugaritic), or by contextualizing the word. Applying our topic modeling architecture to the first road to explanation will be difficult, since the digitization of Aramaic and Ugaritic texts is still in its early steps. But we believe that our method can aid in bringing about a clearer and quantitatively supported discussion of the context of these contested words and their meaning. We believe that topic modeling these words following the manner described above, a new insight into their context can be gained, which is this time motivated by a probabilistic algorithm. This will narrow down the interpretation radius of the semantic field of the words. This is relevant to the discussion of diachronic language evolution, because hapax legomena are usually compared to other extant words believing to have served more or less the same function. Using our topic modeling approach, we can verify until what extent this comparison is justified given our quantitative reasoning, and place it within the diachronic picture we have suggested in the previous paragraph.

Thirdly, we consider the evolution of the Hebrew script. The script makes regular use of so-called matres lectionis (reading mothers), which are consonant signs not read in their consonantal rendering, but used for marking the location of a vowel. As is generally known, Hebrew does not write its vowels, but nevertheless these reading mothers are used to facilitate the process, referring to a set of vowels, never just one. In general it is correct that later writings use these reading mothers more often, resulting in abundant attestation in the Dead Sea Scrolls. What however is still a matter of debate, is the cause of this difference in writing the Hebrew words. Many different suggested explanations exist, among which for example the preservation of the original text including its pronunciation seems important. More interesting for our purposes is the theory that as times progressed, writers were to a lesser extent familiar with the Hebrew language often they were native speakers of Aramaic with a profound knowledge of Hebrew scripture so that they used a more explicit form of writing. A strength of topic modeling is that it can look at semantics uninfluenced by orthographic differences. By comparing the information generated by this approach with the one we get by treating different spellings as different words, we can study whether these differences can be explained by semantic information, which in turn can clarify our insight into orthographic change of the Hebrew language.

References

[1] Franco Moretti. *Graphs, Maps, Trees.* Verso, 2005. 119 p.

[2] Dirk Roorda Wido Van Peursen. Shebanq.