CHAT: A System for Stylistic Classification of Hebrew-Aramaic Texts

Moshe Koppel Dror Mughaz Navot Akiva

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

1. Objectives

CHAT, is a fully self-sufficient system for pre-processing, vectorizing and categorizing Hebrew-Aramaic texts. CHAT is designed to work with Bar-Ilan's corpus of Hebrew-Aramaic texts incorporating over 128 million words spanning more than two millennia. The kinds of problems that are of interest for this system do not concern categorization by topic but rather a number of scholarly issues concerning authorship. In particular, we are concerned with ascertaining answers to the following questions:

1. Authorship Verification: Were two given corpora written/edited by the same author or not?
2. Chronology: Which documents preceded which and did some influence others?
3. Scribal Attribution: From which of multiple versions of the same text might some text fragment have been taken?

These questions lead to some very interesting methodological issues. For example, in answering the first question, we must deal with two types of malicious writers:

a. Forgers – those who deliberately try to fool us into believing that a document was written by a particular author *X*
b. Pseudonyms – Author *X* deliberately tries to fool us into believing that a document was written by some (otherwise unknown) author *not-X*

2. Specific Challenges

We will describe in detail how CHAT solves these problems. Obviously, the two main issues that distinguish CHAT from most other text categorization systems are the stylistic nature of the problems we need to solve and the unique features of Hebrew-Aramaic. Both these issues are significant primarily at the feature selection stage.

Style vs. Content

The last ten years has seen an explosion of research in automated text categorization (Sebastiani 2002). Driven largely by the problem of Internet search, the text categorization literature has dealt primarily with categorization of texts by topic rather than by writing style. There has, however, been a considerable amount of research on authorship attribution. Most of this work has taken place within what is often called the "stylometric" community (Holmes 1998, McEnery & Oakes 2000) which has tended to use statistical methods substantially different in flavor from those typically used by researchers in the machine learning community. Nevertheless, in recent years machine learning techniques have been used with increasing frequency for solving style-based problems. The grand-daddy of such works is that of Mosteller and Wallace (1964) who applied Naïve Bayes to solve the problem of the Federalist Papers. Recent such works include those of Matthews and Merriam (1993, 1994) on the works of Shakespeare, Argamon-Engelson et al (1999) on news stories, Wolters and Kirsten (2000) on genre, de Vel et al (2001) on email authorship, Stamatatos et al (2001) on Greek texts, and Koppel et al (2003) on gender.

Categorization according to topic is a significantly easier problem than categorizing according to author style. The kinds of features which researchers use for categorizing according to topic typically are frequencies of content words. In contrast, for categorizing according to author style one needs to use precisely those linguistic features which are content-independent. In the past researchers have used lexical (Mosteller & Wallace 1964), syntactic (Baayen et al 1996, Argamon-Engelson et al 1999, Stamatatos et al 2001), or complexity-based (Yule 1938) features for this purpose. We will see, though, that there remains a great deal of room for interesting work with regard to choice of feature types.

Language

Hebrew and Hebrew-Aramaic texts present special problems. In particular, function words tend to be conflated into word affixes. Moreover, no parts-of-speech tagger for Hebrew was available to us. Fortunately, though, a great deal of morphological and orthographic information is easily exploitable.

3. Case Studies

We consider three case studies to illustrate specific difficulties that we encountered:

Problem 1: We are given one corpus written by a 19th Century Baghdadi scholar, Ben Ish Chai, and another corpus believed to have been written by him under a pseudonym. We need to determine if the two corpora were written by the same person or not.

Problem 2: We are given three sub-corpora of the classic 12th Century work of Jewish mysticism, Zohar. Scholars are uncertain whether the three corpora were authored by a single author and, if not, which corpora influenced which others. We need to propose the likeliest relationship between the corpora.

Problem 3: We are given four manuscripts of the same tractate of the Babylonian Talmud. The object is to determine from which manuscript a given fragment is taken. (We are given the text of the fragment, not the original, so that hand-writing clues are not relevant.)

The choice of learning algorithm for the actual classification is the least interesting aspect of the work. For the first two problems below, our learning algorithm was a generalization of the Balanced Winnow algorithm of Littlestone (1987) which has previously been shown to be effective for text-categorization by topic (Lewis et al 1996, Dagan et al 1997) and by style (Koppel et al 2003). For the third problem, we needed a learner which could handle more than two classes, so rather than awkwardly adapting Balanced Winnow (which is naturally a binary classifier), we used Naïve Bayes.

4. Experiments and Results

Problem 1: Unmasking Pseudonymous Authors

We are given several collections of responsa (letters written in response to legal queries) written in Hebrew-Aramaic by a number of Iraqi rabbinic scholars in the 18th and 19th centuries. These are:

TL (Torah Lishmah) – 524 documents
RP (Rav Pe'alim) – 509 documents
GV (Ginat Veradim) – 388 documents
SV (Shoel veNishal) – 585 documents
DN (Darhei Noam) – 150 documents

An important historical enigma is associated with *TL*. The author of *RP*, Ben Ish Chai, claims to have found the manuscript of *TL* in an archive. Yet there is ample historical reason to believe that he in fact authored the manuscript but did not wish to take credit for it for personal reasons. What do the texts tell us?

We begin by checking whether we able to distinguish one collection from another using standard text categorization techniques. We select a list of lexical features as follows: the 200 most frequent words in the corpus are selected and all those that are deemed content-words are eliminated manually. We are left with 130 features. Strictly speaking, these are not all function words but rather words which are typical of the legal genre generally, without being correlated with any particular sub-genre. Thus, in the responsa context, a word like *question* would be allowed although in other contexts it would not be considered a function word.

Since the texts we have of 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. Thus, we eliminate all orthographic variations: we expand all abbreviations and unify variant spellings of the same word. After pre-processing the text, we constructed vectors of length 130 in which each element represented the relative frequency (normalized by document length) of each feature.

We then used Balanced Winnow as our learner to distinguish pairwise between the collections. Five-fold cross-validation experiments yield accuracy of greater than 95%

for each pair. In particular, we are able to distinguish between *RP* and *TL* with accuracy of 98.5%. Does this mean that *RP* and *TL* are by different authors?

A cursory glance at the texts indicates that the results do not necessarily support such a conclusion. The author of *TL* simply used certain stock phrases, which do not appear in *RP*, in a very consistent manner. The author of *RP* could easily have written *TL* and used such a ruse to deliberately mask his identity. How can we nail him?

We learned models to distinguish *TL* from each other author. As noted, such models are quite effective. In each case, we then eliminated the five highest-weighted features and then learned a new model. We iterated this procedure ten times. The results (shown in Figure 1) could not be more glaring. For *TL* versus each author other than *RP*, we are able to distinguish with gradually degrading effectiveness as the best features are dropped. But for *TL* versus *RP*, the effectiveness of the models drops right off a shelf. This indicates that just a few features, possibly deliberately inserted as a ruse or possibly a function of slightly differing purposes assigned to the works, distinguish between the works. Once those are eliminated, the works become indistinguishable – a phenomenon which does not occur when we compare *TL* to each of the other collections. Thus we conclude that the author of *RP* probably was the author of *TL*.

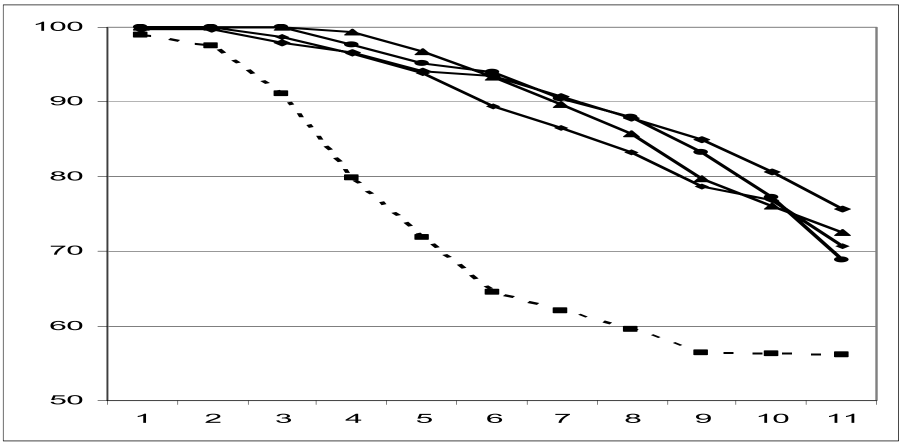

Fig. 1: Accuracy (*y*-axis) on training data of learned models comparing *TL* to other collections as best features are eliminated, five per iteration (*x*-axis). Dotted line on bottom is *RP* vs. *TL*.

Problem 2: Establishing Chronology and Dependence

The corpora used in this experiment were three sub-corpora of the central text of Jewish mysticism:

I: HaIdra (47 passages from Zohar vol. 3, pp. 127b-141a; 287b-296b)
M: Midrash HaNe'elam (67 passages from Zohar vol. 1, pp. 97a-140a)
R: Raya Mehemna (100 passages from Zohar vol. 3, 215b-283a)

As noted above, scholars of Jewish mystical literature (Tishbi, 1949) are uncertain of the provenance of these corpora: were they authored by a single author and, if not, which may have influenced the others.

Lexical features were chosen in a similar fashion to that described above. Experiments were run separately on each pair from among the three corpora. In five-fold cross-validation on each pair, unseen documents were classified with approximately 98% accuracy. Using the methodology described above, we found that these results remained stable even as the best features were eliminated. From this we conclude that these corpora were likely written by three different authors.

The next stage of the experiment is an attempt to determine the relationship between the three corpora. We learn models to distinguish two of the corpora from each other and then use this model to classify the third corpora as more similar to one or the other. In our initial experiments, absolutely nothing could be concluded because in each of the three experiments the passages of the third corpus seemed to split about evenly between being more similar to the first as to the second.

We then ran the experiment again but this time using grammatical prefixes and suffixes as features in addition to the lexical features. We compiled a set of 105 letter bi-grams which can serve as grammatical prefixes (57) and suffixes (48). Since there are many words which begin or end with these bi-grams but in which the bi-grams do not serve as grammatical affixes (but rather are part of the root), we compiled a dictionary of such words which we used as a filter in counting affix frequencies.

Using the expanded feature set, we were able to pair-wise distinguish between the corpora with the same 98% accuracy as with the original lexical feature set. However, the results of the second experiment changed dramatically. When we learn models distinguishing between R and M and then use them to classify I, all I passages are classified as closer to R. Similarly, when we learn models distinguishing between R and I and then use them to classify M, all M passages are classified as closer to R. However, when we learn models distinguishing between M and I and then use them to classify R, the results are ambiguous.

In order to understand why this happens, we need a bit of linguistic background: Like our other corpora, Zohar is written in a dialect that combines Aramaic and Hebrew. One of the main distinguishing features of Hebrew versus Aramaic is the use of certain affixes. For example, in Hebrew the plural noun suffixes are ות and ים, while in Aramaic ין and נא are used. Similarly, in Hebrew *which* is incorporated as the prefix ש while in Aramaic ד is used. We find that M is characterized by a large proportion of Hebrew affixes and I is distinguished by a large proportion of Aramaic affixes. R falls neatly in the middle.

A number of possible conclusions might be drawn from this. For example, the phenomena uncovered here might support the hypothesis that R lies chronologically between M and I. However, scholars of this material believe that a more likely interpretation is that M and I were co-temporaneous and independent of each other and that R was subsequent to both and may have drawn from each of them.

Problem 3: Assigning Manuscript Fragments

We are given four versions of the same Talmudic text (tractate Rosh Hashana of the Babylonian Talmud), each version having been transcribed by a different scribe. We break each of the four manuscripts into 67 fragments (corresponding to pages in the printed version). The object is do determine from which version a given fragment might have come.

Note that since we are distinguishing between different versions of the same texts, we can't realistically expect lexical or morphological features to distinguish very well. After all, the texts consist of the same words. Rather, the features which are likely to help here are precisely those which were disqualified in our earlier experiments, namely, orthographic ones.

Rather than identify these features manually, we proceeded as follows. First, we simply gathered a list of all lexical features which appeared at least ten times in the texts. Variant spellings of the same word were treated as separate features. In order to identify promising features, we used an entropy measure which grants a high score to a feature which appears with different frequency in different versions of the same document.

Specifically, let $\{d_1,d_2,\ldots,d_n\}$ be a set of texts (in our case n=67) and let $\{d_i^1,d_i^2,\ldots,d_i^m\}$ be m > 1 different versions of d_i (in our case m=4). For each feature c, let c_i^j be the relative frequency of c in document d_i^j. For multiple versions of a single text d_i, let $k_i = \Sigma_j c_i^j$ and let $H(c_i) = -\Sigma_j [(c_i^j/ k_i)\log (c_i^j/ k_i)]]/\log m$. (We can think of c_i^j/ k_i as the probability that a random appearance of c in d_i is in version d_i^j so that $H(c_i)$ is just the usual entropy measure.) Thus, for example, if a feature c assumed the identical value in every version of a document d_i, $H(c_i)$ would be 1. To extend the definition to the whole set $\{d_1,d_2,\ldots,d_n\}$, let $K = \Sigma_i k_i$ and let $H(c) = \Sigma_i [(k_i/K) * H(c_i)]$. Finally, let $H'(c) = 1-H(c)$. $H'(c)$ does exactly what we want: features the frequency of which varies in different versions of the same document score higher than those which have the same frequency in each version.

We then ranked all features according to $H'(c)$. Those which ranked highest were those which permitted variant orthographic representations. In particular, some scribes used abbreviations or acronyms or non-standard spellings in places where other scribes did not. We choose as our feature set the 200 highest-ranked features according to H'. Using Naïve Bayes on this feature set in five-fold cross-validation experiments yielded accuracy of 85.4%.

Thus, by and large, we are able to correctly assign a fragments with its manuscript of origin. This work recapitulates and extends in automated fashion, a significant amount of research carried out manually by scholars of Talmudic literature (Friedman 1996).

There is one major limitation to the approach we used here. We assume that within a given manuscript the frequency of a given feature is reasonably invariant from fragment to fragment. This is only true if we are considering various versions of a single thematically homogeneous text. If we wish to train on versions of various texts as a basis for identifying the scribe/editor of a manuscript of a different text, we need make a more realistic assumption. This can be done by normalizing our feature frequencies differently: we must count the number of appearances of a particular orthographic variant of a word in a manuscript fragment relative to the total number of appearances of all variants of that word in the fragment. This value should indeed remain reasonably constant for a single scribe/editor across all texts.

5. Conclusions

We have shown that the range of issues considered in the field of text categorization can be significantly broadened to include problems of great importance to scholars in the humanities. Methods already used in text categorization require a bit of adaptation to handle these problems. First, the proper choice of feature sets (lexical, morphological and orthographic) is required. In addition, juxtaposition of a variety of categorization experiments can be used to handle issues of pseudonymous writing and chronology in surprising ways. We have seen that for a variety of textual problems concerning Hebrew-Aramaic texts, proper selection of feature sets combined with these new techniques can yield results of great use to scholars in these areas.

6. References

Argamon-Engelson, S., M. Koppel, G. Avneri (1998). Style-based text categorization: What newspaper am I reading?, in Proc. of AAAI Workshop on Learning for Text Categorization, 1998, pp. 1-4

Baayen, H., H. van Halteren, F. Tweedie (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution, Literary and Linguistic Computing, 11, 1996.

Dagan, I., Y. Karov, D. Roth (1997), Mistake-driven learning in text categorization, in EMNLP-97: 2nd Conf. on Empirical Methods in Natural Language Processing, 1997, pp. 55-63.

de Vel, O., A. Anderson, M. Corney and George M. Mohay (2001). Mining e-mail content for author identification forensics. SIGMOD Record 30(4), pp. 55-64

Friedman, S. (1996) The Manuscripts of the Babylonian Talmud: A Typology Based Upon Orthographic and Linguistic Features. In: Bar-Asher, M. (ed.) *Studies in Hebrew and Jewish Languages Presented to Shelomo Morag* [in Hebrew], p. 163-190. Jerusalem, 1996.

Holmes, D. (1998). The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing, 13, 3, 1998, pp. 111-117.

Koppel, M., S. Argamon, A. Shimony (2003). Automatically categorizing written texts by author gender, Literary and Linguistic computing, to appear

Lewis, D., R. Schapire, J. Callan, R. Papka (1996). Training algorithms for text classifiers, in Proc. 19th ACM/SIGIR Conf. on R&D in IR, 1996, pp. 306-298.

Littlestone, N. (1987). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, Machine Learning, 2, 4, 1987, pp. 285-318.

Matthews, R. and Merriam, T. (1993). Neural computation in stylometry : An application to the works of Shakespeare and Fletcher. Literary and Linguistic computing, 8(4):203-209.

McEnery, A., M. Oakes (2000). Authorship studies/textual statistics, in R. Dale, H. Moisl, H. Somers eds., Handbook of Natural Language Processing (Marcel Dekker, 2000).

Merriam, T. and Matthews, R. (1994). Neural computation in stylometry : An application to the works of Shakespeare and Marlowe. Literary and Linguistic computing, 9(1):1-6.

Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist.* Reading, Mass. : Addison Wesley, 1964.

Sebastiani, F. (2002). Machine learning in automated text categorization, ACM Computing Surveys 34 (1), pp. 1-45

Stamatatos, E., N. Fakotakis & G. Kokkinakis, (2001). Computer-based authorship attribution without lexical measures, Computers and the Humanities 35, pp. 193—214.

Tishbi, Y. (1949). *Mishnat haZohar* (in Hebrew), Magnes: Jerusalem, 1949.