### The literature case study

Assume we have two sets of texts. The first set is several books of historical accounts where each account is preceded by a chain of narrators who narrated the account. The second set is several books of biographies where each biography of a person may include an evaluation of the authenticity of that person. We want a tool that can answer a complex query such as relating all historical accounts in the first set that discuss women and vehicles (such as tamed animals) to the narrators of the accounts, and then relate these narrators to the second set of books of biographies with an emphasis on authenticity evaluations.

We need tools that can search and index the historical accounts to return a subset that relates to women and travel. Then we need tools that can identify names, and succession relations between names to build a chain of narrators for each historical account in the subset. We need tools that can identify the names in the biographies and detect evaluation adjectives.

Provided we have these tools we can exhaustively at the click of a button visualize the authenticity of all the historical accounts that may allow or forbid women from driving according to Islamic laws. We can go one step further to detect locations and dates and relations based on chronological orders and further question the consistency of the historical accounts. Such query takes currently huge amounts of manual effort that is also prone for error and far from being exhaustive.

We here envision that text mining applied to the authenticity of historical accounts will have a huge impact in the field, especially when these historical accounts under consideration are the “hadith” (حديث). A hadith in Islamic literature is a narration from the prophet Mohammad related by multiple narrators. Establishing the authenticity of a hadith is an important task in Islamic studies.[[1]](#footnote-1) Several researchers ~\cite{Hadithopaedia:08} attempted to automate the analysis of the hadith literature.

Figure 1

A hadith in Islamic literature is a narration from the prophet Mohammad related by multiple narrators.

Establishing the authenticity of a hadith is an important task

in Islamic studies.



Figure 1: Literature Case-study representation

%Several researchers~\cite{Hadithopaedia:08} attempted to automate

%the analysis of the hadith literature.

In this paper we use Sarf to successfully automate the

analysis of three books of hadith selected

arbitrarily~\cite{IbnHanbal,AlKulayni,AlTousi}\footnote{We obtained

the digitized books from online sources such as

\href{http://www.yasoob.com/}{http://www.yasoob.com/} and

\href{http://www.al-eman.com/}{http://www.al-eman.com/}. }

% Our understanding is that those books are historical documents with

% public IP.}.

%\setcode{utf8}

%\begin{arabtext}

%Ø­Ø¯Ø«Ù†Ø§ \emphasize{Ù‡Ø´ÙŠÙ…}ØŒ Ø£Ø®Ø¨Ø±Ù†Ø§ \emphasize{Ø¥Ø³Ù…Ø§Ø¹ÙŠÙ„ Ø¨Ù† Ø£Ø¨ÙŠ Ø®Ø§Ù„Ø¯}ØŒ

% Ø¹Ù† Ø£Ø¨ÙŠ Ø¥Ø³Ø­Ø§Ù‚ØŒ Ø¹Ù† Ø³Ø¹ÙŠØ¯ Ø¨Ù† Ø¬Ø¨ÙŠØ±ØŒ Ù‚Ø§Ù„ ÙƒÙ†Ø§

% Ù…Ø¹ Ø§Ø¨Ù† Ø¹Ù…Ø± Ø±Ø¶ÙŠ Ø§Ù„Ù„Ù‡ Ø¹Ù†Ù‡ Ø­ÙŠØ« Ø£ÙØ§Ø¶ Ù…Ù†

% Ø¹Ø±ÙØ§Øª Ø¥Ù„Ù‰ Ø¬Ù…Ø¹ ÙØµÙ„Ù‰ Ø¨Ù†Ø§ Ø§Ù„Ù…ØºØ±Ø¨ ÙˆÙ…Ø¶Ù‰

%\end{arabtext}

%\setcode{standard}

Our analysis accepts a book as input

and segments it into a vector of hadiths.

We segment each individual hadith into two parts.

The first part is the sanad and it

contains the chain of narrators who related the hadith

from Mohammad.

The second part is the matn or the content

of the hadith.

We are concerned to further explore the sanad and

we detect the chain of narrators and

the relation that links each narrator to his ancestor and

his predecessor in the chain.

We also detect the full name of each narrator that is

usually composed of several proper names with connectors

in between.

We built our hadith case-based controller to target

the detection of proper names.

This includes the task of finding compound names

composed of several words such as \RL{`bd alr.hmn} often

appearing as ``run-on'' words.

The controller also targets words that mean narrate when

they appear in the neighborhood of multiple proper names.

\subsection{Controller}

\label{sec:controller}

% FSM for the controller

We illustrate the hadith case-based controller FSA

in Figure~\ref{f:hadith}.

The transitions in the hadith FSA are excited

by inputs from the Sarf morphological analyzer.

A NAME label means that Sarf encountered a proper name.

An NRC label means that Sarf met a narrator connector such as

\RL{`an} "on behalf of", \RL{.hada\_t} "narrated", \RL{qaal} "said",

\RL{'a\_hbar} "told" or one of their derivations.

The IBN label corresponds to the word \RL{ibn} "the son of" and is a name connector.

The NISBA label corresponds to an adjective that points to a person such

as \RL{alma.sriy} "the Egyptian".

The symbol $\tau\_{\mbox{NMC}}$ is a threshold

that corresponds to the number of tolerated name connectors

that may occur between two names.

The symbol LIST$\_{\mbox{NMC}}$ corresponds to the list

of name connectors collected since the controller

started looping in the state NMC\\_S.

The symbol $\lambda\_{\mbox{NMC}}$ is a parameter

that corresponds to a relaxed tolerance measure that

the controller resorts to in case the words separating

two names were longer than $\tau\_{\mbox{NMC}}$ but

contained a name connector word such as IBN or NISBA.

The controller has four states that correspond to

an abstract position in text relative to the next sanad.

State TEXT\\_S is the initial state and denotes that

the controller is outside the context of a sanad.

The controller moves to the state NAME\\_S whenever

NAME is reported by Sarf.

State NRC\\_S means that the controller thinks it is in the context

of a narrator name after Sarf reports an NRC.

State NMC\\_S

indicates that the controller expects a name to appear within

a tolerance threshold expressed by

$\tau\_{\mbox{NMC}}$ and $\lambda\_{\mbox{NMC}}$.

\begin{figure}[tb!]

\center{

\resizebox{.9\columnwidth}{!}

{ \input{psfigs/hadith.pstex\_t}}

\caption{FSA controller of the hadith case study.}

\label{f:hadith}

}

\end{figure}

Similarly, the NRC\\_S state tolerates $\tau\_{\mbox{NRC}}$ words

before it gives up on its expectations.

We reach the NAME\\_S state only when a

valid NAME is detected and we leave when no more NAME's are detected.

The NRC\\_S state can only be reached if an NRC is detected.

NMC\\_S can only be reached from a NAME\\_S state.

The FSA reports a valid chain of narrators when a sequence of names

connected by narrator connectors appears.

It marks the beginning of the hadith with the beginning of the current sequence,

and marks the end of the hadith with the beginning of the next sequence.

It also marks the chain of narrators as the sequence itself.

The definition of input labels such as NAME and IBN depends on the

morphological analyzer.

However, our case-based controller approach was able to perform well

under both Sarf and a refined version of Sarf.

## References (reproduced here for convenience)

x

|  |  |
| --- | --- |
| [1] | J. Nightingale, "Digging for data that can change our world," *The Guardian*, 10 January, 2006. |
| [2] | A. Srivastavaand M. Sahami, *Text Mining: Classification, Clustering, and Applications*.: Chapman & Hall, 2009. |
| [3] | H. Al-Fawareh and S. Jusoh and W. Sheikh Osman, "Ambiguity in Text mining," *IEEE proceedings of the international conference on computer and communication engineering*, May 2008. |
| [4] | J. Han and M. Kamber, *Data mining concepts and techniques*.: University of Illinois at Urbana Champaign, 2005. |
| [5] | K Beesley, T Buckwalter, and S Newton, "Two-level finite-state analysis of Arabic morphology," in *Seminar on Bilingual Computing in Arabic and English*, Cambridge, England, 1989. |
| [6] | Tim Buckwalter, "Buckwalter Arabic morphological analyzer version 2.0," Linguistic Data Consortium (LDC) , catalog number LDC2004L02, 2004. |
| [7] | (2009) Columbia's Arabic Dialect Modeling Group. [Online]. <http://www1.ccls.columbia.edu/~cadim/> |
| [8] | A. Al-Zoghby and A. SharafEldin and N. Ismail and T. Hamza, "Mining Arabic Text Using Soft-Matching," in *Computer Engineering & Systems*, 2007. |
| [9] | A. EL-Halees, "Arabic Text Classification. Using Maximum Entropy," *Islamic University journal* , 2007. |
| [10] | Mesleh Abdelwadood, "Support vector machines based Arabic language text classification system: feature selection comparative study," in *12th WSEAS International Conference on Applied Mathematics*, 2007. |
| [11] | M Eldost, "Arabic text data mining: A root-based hierarchical indexing model," *International journal of modelling & simulation*, vol. 23, 2003. |
| [12] | Abdelhadi Soudi, Antal van den Bosch, and Neumann Guenter, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*.: Springer, 2007. |
| [13] | (2009, September) Sakhr Inc. [Online]. <http://www.sakhr.com/products/Mining> |
| [14] | (2009) Basis technology. [Online]. <http://www.basistech.com/base-linguistics/arabic/> |
| [15] | Fadi Zaraket, Rabih Zbib, and Monah Khalil, "Yaraah: Arabization module for Windows NT," AUB, technical FYP. 1996. |
| [16] | Fadi Zaraket, Marwan Zeineddine, Hussein AlAsadi, and et al. (2006) Hadithopedia: comprehensive hadith encylopedia. [Online]. <http://code.google.com/p/hadithopaedia/> |
| [17] | Judy Redfearn, "Text Mining," *Joint Information Systems Committee*, September 2008. |
| [18] | R. Shartouni, *The principles of Arabic/Mabādi al-Arabiyah*. Beirut: Le Levant, 1973. |
| [19] | Abdeh AlRajehi, *Semantical Practice/At-tatbiq Alnahawi*, 1st ed. Beirut: Renaissance (nahda), 2000. |
| [20] | Abdeh AlRajehi, *The morphological practice/At-tatbiq Assarfi*, 1st ed. Beirut: Renaissance (An-nahda), 2000. |
| [21] | Ryan, Owen Rambow, Nizar Habash, Mona Diab, and Cynthia Rudin Roth, "Arabic Morphological Tagging, Diacritization, and Lemmatization Using Lexeme Models and Feature Ranking," in *Association for Computational Linguistics (ACL)*, Columbus, Ohio, 2008. |
| [22] | G. Kanaan, M. Yaseen, B. Al-Sarayreh, N. A. Al-Naji R. Al-Shalabi, "Arabic Query Expansion Using Interactive Word Sense Disambiguation," in *International Conference on Arabic Language Resources & Tools*, cairo, 2009. |
| [23] | (2009) International Journal of Arabic-English Studies. [Online]. <http://www.ijaes.org/> |
| [24] | (2009) DokuWiki. [Online]. <http://www.dokuwiki.org> |
| [25] | (2009) Subversion. [Online]. <http://subversion.tigris.org/> |
| [26] | Niklas Een and Niklas Sörensson, "An Extensible SAT-solver," in *SAT*, 2003. |
| [27] | QT C++ Library. [Online]. <http://www.qtsoftware.com> |

x

1. Note that most of the Islamic history and Hadith books as well as a huge library of Arabic literature books are already digitized and available on the internet. [↑](#footnote-ref-1)