School of Computing
FACULTY OF ENGINEERING

UNIVERSITY OF LEEDS

Using the Web to Model Modern and Quranic Arabic

Eric Atwell

Language Research Group

I-AIBS: Institute for Artificial Intelligence and Biological Systems

School of Computing

University of LEEDS, England

Overview

UNIVERSITY OF LEEDS

Artificial Intelligence and Corpus Linguistics at Leeds Uni
Using the Web to Model Modern and Quranic Arabic

Web-based software and corpus datasets from Leeds:

Modern Standard Arabic and Quranic Arabic

Interest: not only Arabic corpus/computational linguists; also Quranic students, and the general public.

Proposals for further work: the Quranic Knowledge Map;

LREQ: Language Resources and Evaluation and the Quran

Artificial Intelligence and Corpus Linguistics

UNIVERSITY OF LEEDS

Corpus: a collection of text, representing a topic or task

Corpus Linguistics: study of language based on a Corpus

AI: Machine Learning “learns” patterns, rules from data

Text Analytics: ML “useful” patterns from text data

Example research using ML to learn from a corpus: ...

Classifying Cause of Death in Verbal Autopsies

UNIVERSITY OF LEEDS

Verbal Autopsy: interview of a mother after her baby died

e.g. In Ghana, to gather WHO stats on Causes of Death

10,000 VAs sent to London LSHTM, doctors diagnose CoD

ML: learn patterns linking features of each VA to CoD

- To predict CoD in future VAs, without need for doctors
- To guide health funding policy, NOT front-line health care

(funded by Association of Commonwealth Universities)

Predicting prosody: when to pause while reading a text

UNIVERSITY OF LEEDS

When you read a text, you pause at commas, full-stops, ...

... Pauses can also be natural at other places

... eg in text without punctuation: poetry, web-text etc.

ML from a corpus of text read out loud: BBC radio broadcasts

To predict phrase breaks in Text-to-Speech

(may also apply to classical Arabic poetry, Quran, ...?)

Making Sense

UNIVERSITY OF LEEDS

Goal: to develop systems to better manage data collected in connection with alleged terrorist plots.

“like looking for a needle in a haystack.”

I prefer the analogy of looking for threads in a haystack

ML to find “interesting” texts, and “threads” linking them

Needs a training corpus, where “interesting” texts are marked

(funded by UK EPSRC, ESRC, CPNI)

CPNI

Centre for the Protection of National Infrastructure

ML from data

UNIVERSITY OF LEEDS

Confession: I am NOT an Arabic linguist!

So, how can I be involved in Arabic corpus linguistics research?

Machine Learning requires analysis of data to extract features and patterns – I do not have to "understand" the data

I am NOT:

- A doctor – but maybe ML can help classify CoD from VAs
- A counter-terrorism expert – but maybe ML can help detect terrorist threads in data

Using the Web to Model Modern and Quranic Arabic

UNIVERSITY OF LEEDS

Using the Web:

... as source of corpus data

- Scouting for websites with “good” data
- BootCat: automate harvesting of web-page text

... to publicise and promote re-use of Corpora

- put corpora and tools on WWW, open-source

... to annotate corpus: “crowd sourcing”

- volunteers can build a shared resource

Using the Web to Model Modern and Quranic Arabic

UNIVERSITY OF LEEDS

... **to model** ...

Computational Modelling

- use corpus as “training data” for Machine Learning

Linguistic theories or models

- eg traditional Arabic grammar can be modelled: Treebank
- eg morphology model applied to Arabic Web Corpus

Arabic computing research at Leeds: Modern Arabic

UNIVERSITY OF LEEDS

Abc – Arabic by computer: online texts for language students

Arabic corpus-trained chatbot

Corpus of Contemporary Arabic

aConCorde: concordance for Arabic texts

SALMA morphological analysis and tag-set

Arabic lexical resource from traditional Arabic dictionaries

Discourse Treebank for Modern Standard Arabic

180-Million-word Arabic Web Corpus, online concordance

http://www.comp.leeds.ac.uk/arabic

Arabic computing research at Leeds: Quran as Corpus

UNIVERSITY OF LEEDS

Quran chatbot: replies with verse from Quran

Qurany: browse Quran by concepts

Morphochallenge: Quran as Gold Standard for evaluation

Quranic Arabic Corpus: morphology and syntax annotations

Text mining the Quran: related verses; pronoun coreferences

(Web-as-Corpus approach to populating Wikiversity for teaching about Islam and Muslims)

http://www.comp.leeds.ac.uk/arabic

http://www.comp.leeds.ac.uk/arabic

UNIVERSITY OF LEEDS

Latifa Al-Sulaiti has developed a new free-to-download Arabic corpus, the **Corpus of Contemporary Arabic**

Andy Roberts has developed open-source concordance tool for analysis of Arabic corpus texts, **aConCorde**

Majdi Sawalha has developed an Arabic morphological analysis tool to extract **Arabic word root**

Nora Abbas has developed a Quran "search for a concept" tool and website, **Qurany**

Kais Dukes is developing an online annotated linguistic resource which shows grammar, syntax and morphology for each word in the Holy Quran, the **Quranic Arabic Corpus**

AbdulBaquee Sharaf – Text Mining The Quran

Wordle of the Quran

UNIVERSITY OF LEEDS

Wordle after correction

UNIVERSITY OF LEEDS

Open-source and on the WWW

UNIVERSITY OF LEEDS

Our resources are open-source rather than commercial; this is why they have been widely re-used, compared to resources kept “in-house” by other Arabic NLP research groups.

Our Quranic Arabic Corpus website http://corpus.quran.com/ shows the advantages of making resources open-source : publications, press articles, Message Board for feedback, Google Analytics visualisation of global distribution of visitors to the website.

Understanding the Quran – a Grand Challenge for AI

UNIVERSITY OF LEEDS

Understanding Islam is a major societal issue:

- Western schools, universities and the general public need an objective, impartial online Quran Expert to learn about Islam
- non-Arabic-speaking Muslims may also be ignorant of the deeper meanings in the Quran, despite memorising recitation

Understanding the Quran – a Grand Challenge for AI

UNIVERSITY OF LEEDS

Current systems can search for words, and fact questions eg "are angels male?" ... But we need a new Knowledge Representation and Reasoning formalism capable of capturing complex, subtle knowledge encoded in the Quran

Understanding the Quran – a Grand Challenge for AI

UNIVERSITY OF LEEDS

Machine Learning research needs a “Gold Standard” – a corpus where each text is classified and marked up by experts, so ML can learn the classification.

(for Making Sense, we need a Gold Standard where some texts are marked by experts as “interesting”)

The Quran is an excellent Gold Standard: many expert analyses exist (Tafsir), we can use these to train ML

Quranic scholarly work can ensure that Knowledge Based Systems based on the Quran are logically consistent and correct

Understanding the Quran – a Grand Challenge for AI

UNIVERSITY OF LEEDS

Huge worldwide interest in the Quran means we can harness volunteers for “crowd-sourcing” analysis

Quranic Arabic Corpus: initial automatic analysis, then proofreading and correction by many volunteers

A proposal for further research: the Quranic Knowledge Map

Understanding the Quran is a grand challenge for society, for western public education, for Muslim-world education, for knowledge representation and reasoning, for knowledge extraction from text, for systems robustness and correctness, and for online collaboration.

Understanding the Quran is a grand challenge for computer science and artificial intelligence

We propose a collaborative research effort to construct a Quranic Knowledge Map to address this challenge.

Three strands of research

UNIVERSITY OF LEEDS

Infrastructure. A set of tools used to develop the Quranic Knowledge Map: Arabic Natural Language Processing tools, tools for online collaborative annotation, and tools for knowledge engineering and automated reasoning.

Datasets. Tagging the Quran with morphology, syntax, semantics, pronoun and named entity references, concept ontology, other KR formalisms. Also, extending beyond Quran to linked Classical Arabic texts: Hadith etc. Each of these datasets is expected to be highly useful for further research and worthy in publication and distribution in itself.

End-user applications. These form the main contribution of the Quranic Knowledge Map to society, i.e. to interested researchers, students and public who will use the system.

Modules in the Quranic Knowledge Map

UNIVERSITY OF LEEDS

	INFRASTRUCTURE	DATASETS	APPLICATIONS
	Reusable Computational Tools	The Quranic Arabic Corpus	Online User Access
CORE TEXT ANNOTATION	Software for Classical Arabic NLP Arabic Morphological Analyzer Arabic Syntactic Parser Arabic Natural Language Toolkit Multi-lingual Word Alignment	Quranic NLP Datasets Quranic Arabic Text Morphological Tagging Syntactic Treebank Translations & Audio	Baseline Quranic Resource Online Tagged Quran Morphological Search Quranic Grammar Annotations Interlinear Translations
FURTHER LINGUISTICS	Tools for Collaborative Annotation Linguistic Database Manual Annotation Tools Online Collaborative Annotation	Quranic Linguistic Datasets Pronoun Resolution Named Entity Resolution Quranic WordNet	Quranic & Arabic Linguistics Electronic Lexicon & Dictionary Word-sense Disambiguation Arabic Educational Resources
QURANIC KNOWLEDGE	Knowledge Representation Semantic Annotation Framework Semantic Database QA & Information Retrieval Automated Reasoning and Inference	Quranic Knowledge Datasets Ontology of Quranic Concepts Quranic PropBank / FrameNet Related Texts: The Hadith Knowledge Representation	Quranic Knowledge Online Concept Topic Map & Search Verse similarity concordance Quran to Hadith Linkage Question Answering

Research Work-Packages

UNIVERSITY OF LEEDS

WP1 Project Management

WP2 Design:

2.1 User requirements analysis

2.2 Design and specification

WP3: Implementation

3.1 Online collaboration framework

3.2 Morphological, syntactic and semantic taggers

3.3 Tagset design: morphosyntactic dependency and semantic tags

3.4 Interaction and visualization

3.5 Adding other related texts

WP4: Annotation: tagging and proofreading

WP5: Validation and User Evaluation: Case Studies

WP6: exploring applications in Artificial Intelligence research

6.1 Machine Learning of annotations, to tag other related texts

6.2 Learning similarity, links and bridging

WP7: e-learning customization

Collaborators needed

UNIVERSITY OF LEEDS

Researchers with the following expertise:

Management and coordination: leading researchers acting in a supervisory capacity, ideally with skills in fund-raising and public relations!

Religious Studies experts, with experience in the Quran and exegesis (tafsir); and also experts in other religious texts

Full time annotators, familiar with Arabic and the Quran

NLP people, with good proven Arabic language computation skills

Software engineers for general infrastructure and web development, not necessarily with NLP skills

E-learning experts, ideally with a background in developing Arabic online language-learning or religious educational resources

In addition, it is expected that the project will leverage a large body of existing expert volunteers worldwide through collaborative annotation

LREQ

UNIVERSITY OF LEEDS

?- a follow-on workshop for Quran Corpus Linguistics

Language Resources and Evaluation Conference

LREC'2012 Istanbul

?- pre-conference workshop:

?- Language Resources and Evaluation and the Quran

LREQ

Conclusions

UNIVERSITY OF LEEDS

Artificial Intelligence and Corpus Linguistics can apply Machine Learning to "learn" useful patterns in data

Open-source software and corpus datasets:

Modern Standard Arabic and Quranic Arabic

Interest, not only from Arabic computational linguists, but also from Quranic students, and the general public.

Proposal for further work: the Quranic Knowledge Map.