QurSim: A corpus for evaluation of relatedness in short texts

Abdul-Baquee M. Sharaf, Eric S. Atwell
School of Computing
University of Leeds
Leeds, LS2 9JT
United Kingdom
E-mail: a.m.sharaf08@leeds.ac.uk, e.s.atwell@leeds.ac.uk

Abstract

This paper presents a large corpus created from the original Quranic text, where semantically similar or related verses are linked together. This corpus will be a valuable evaluation resource for computational linguists investigating similarity and relatedness in short texts. Furthermore, this dataset can be used for evaluation of paraphrase analysis and machine translation tasks. Our dataset is characterised by: (1) superior quality of relatedness assignment; as we have incorporated relations marked by well-known domain experts, this dataset could thus be considered a gold standard corpus for various evaluation tasks, (2) the size of our dataset; over 7,600 pairs of related verses are collected from scholarly sources with several levels of degree of relatedness. This dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs. This dataset was incorporated into online query pages where users can visualize for a given verse a network of all directly and indirectly related verses. Empirical experiments showed that only 33% of related pairs shared root words, emphasising the need to go beyond common lexical matching methods, and incorporate -in addition- semantic, domain knowledge, and other corpus-based approaches.

Keywords: Text relatedness, Quran, Information Retrieval

1. Introduction

1.1 Text Relatedness

The ability to quantify computationally semantic relatedness of natural language short texts has many interesting application such as: words sense disambiguation, information extraction and retrieval, automatic indexing, lexical selection, text summarization, automatic correction of word errors, and word and text clustering. Although this task is complex computationally, humans routinely perform semantic relatedness tasks readily both at word level, e.g., between words "*cat*" and "*mouse*", or at phrase and text level, e.g., between "*drafting a letter*" and "*writing an email message*". This task has been fairly natural for humans because they can associate a huge amount of background experience and external domain concepts, whereas computational methods lack this smart association mechanism with related external sources.

This paper describes a corpus of short texts marked with relatedness information judged by human domain expert. The degree of relatedness between texts in this corpus varies greatly: although there are instances where lexical matching is evident between the terms in a pair of related texts, the majority of instances require deep semantic analysis and domain specific world knowledge in order to relate the two texts in the pair.

Our objective in collecting this dataset is to provide evaluation and training – and perhaps gold-standard-resource for researchers in the field of computational semantic similarity and relatedness analysis in natural language texts. We chose the Quran for this purpose as it portrays the diverse ways in which scattered verses are related. More elaboration on this is given in the next sub-section.

1.2 Why this text?

The Quran is a scripture which is according to Muslims verbatim words of Allah containing over 77,000 words revealed through Archangel Gabriel to Prophet Muhammad over 23 years beginning in 610 CE. It is divided into 114 chapters of varying sizes, where each chapter is divided into verses, adding up to a total of 6,243 verses.

We chose the Quran for creation of this language evaluation resource for several reasons:

1. Distribution of a particular concept or subject over many scattered verses within different chapters is very evident in the Quran. Often a concept summarized in one verse is elaborated in another verse. Historical events, stories of prophets, emphasis on a command, attributes and qualities of Gods, description of paradise and hell fire, are some of the common subjects that are often repeated in the Quran. However, each repetition adds new meanings absent in other instances, and the overall subject could be fully understood when all instances are taken into consideration. This property of the Quran made it an attractive text for the purpose of analyzing semantic relatedness between short texts, which in our case are between individual verses, or a group of verses of the Quran. We will discuss some types of relatedness in the next section.

2. The Quranic scripture is a widely used and cited document that guides the lives of over 1.5 Muslim adherents today. Increasingly non-Arabic speaking Muslims –and many non-Muslims- learn classical Arabic with the objective of understanding the Quran. For Arabic speakers, the Quran is considered to be the finest piece of literature in the Arabic Language. Producing language evaluation resources for computational analysis of such an important text should be well justified. Moreover, there have been a number of language resources around other spiritual scriptures like the Bible (Resnik et al., 1999).

3. Classical Arabic text is the form of Arabic language used in literary texts authored by early Arabic scholars mainly in the 6th through 10th century. The Quran is considered to be the highest form of classical Arabic text and has been extensively analyzed and cited in linguistic scholarly works since the 7th century. In contrast to other languages, the total body of Arabic texts published during this classical period is as large as modern corpora of Modern Standard Arabic (MSA) - the form used in contemporary scholarly published works as well as in the media- and other modern languages. MSA does not differ from Classical Arabic in morphology or syntax, but richness of stylistic and lexical usage is apparent in Classical works. This makes Classical Arabic subsume MSA making computational and linguistic research work on CA benefit both. However, most recent work on Arabic corpus annotation concentrated on MSA, as a result barring the computational corpus linguistic community from efficient study of a large body of materials available in Classical Arabic.

4. Being a central text in Arabic, over the past 14 centuries a huge body of scholarly commentary volumes has been compiled elaborating on linguistic, stylistic, semantic and other aspects of the Quran. This makes the task of compiling evaluation datasets and annotated corpora on the Quran simpler; as it is usually possible to find scholarly comments on any difficult annotation question. In our QurSim related-verse dataset, we relied on the Quranic commentary work of Ibn Kathir -a well-known Quranic scholar who died in 1373 CE. More elaboration on this text in given in section 2.2.

5. The Quran is widely available translated into almost all live languages of the world, and in many cases there are multiple translations within one language. Among these translations a number of them are also available in machine readable electronic format in the web [1]. All translations maintain chapter and verse numbers as available in the original text, allowing alignment between these translations at equivalent to sentence level. Moreover, to be faithful to the words of God, all translations are made very carefully. Given this fact, any language evaluation resource in the source language of the Quran could be used to evaluate computational tasks on the target language translations as well.

6. The size of the Quran is manageable for manual or semi-automatic annotation tasks. Given that the Arabic language still lacks many NLP resources available for a language like English (e.g., taggers, parsers, Wordnet, frameNet, etc.), developing a manually annotated language resource on a small scale like the Quran text could be a good starting point. A case in point is the Quranic Arabic Corpus (QAC) [2] project (Dukes et al 2010), where every word of the Quran is tagged with morphological, part-of-speech and syntactic information, and is publicly available for research purposes. Another available resource is QurAna, pronominal anaphora of the entire Quran tagged with pronoun antecedent references (Sharaf & Atwell, 2012)[3]. Our QurSim dataset, along with these other available resources on the Quran, will enable interesting computational linguistic applications on the Quran which in turn will eventually motivate wider applications and resource development for Classical and Modern Standard Arabic.

1.3 Text relatedness in the Quran

The Quran has testified that God has deliberately split information within this book and that its verses are semantically paired. Verses 17:106 and 39:23 say [4] respectively:

وَقُرْآنًا فَرَقْنَاهُ لِتَقْرَأَهُ عَلَى النَّاسِ عَلَىٰ مُكْثٍ وَنَزَّلْنَاهُ تَنزِيلًا

And the Quran / We have divided, / that you might recite it / to / the people / at / intervals. / And We have revealed it / (in) stages. /

And [it is] a Qur'an which We have separated [by intervals] that you might recite it to the people over a prolonged period. And We have sent it down progressively.

اللَّهُ نَزَّلَ أَحْسَنَ الْحَدِيثِ كِتَابًا مُّتَشَابِهًا

Allah / has revealed / (the) best / (of) [the] statement - / a Book / (its parts) resembling each other / oft-repeated. /

Allah hath (now) revealed the fairest of statements, a Scripture consistent, (wherein promises of reward are) paired (with threats of punishment)..

In what follows, we discuss some examples to illustrate various forms of relatedness among Quranic verses.

1. Both the following two related verses describe the fate of mountains at the end of times:

وَسُيِّرَتِ الْجِبَالُ فَكَانَتْ سَرَابًا

And are moved / the mountains / and become / a mirage. /

And the hills are set in motion and become as a mirage. [78:20]

وَتَكُونُ الْجِبَالُ كَالْعِهْنِ الْمَنفُوشِ

And will be / the mountains / like wool, / fluffed up/

And the mountains will become as carded wool.[101:5]

2. In the following pair of verses, the first verse uses an Arabic verb that could bear two meanings (arrive, or depart) and the second verse resolves the ambiguity (author's addition within square brackets):

وَاللَّيْلِ إِذَا عَسْعَسَ

And the night / when / it departs, /

And the close [arrival or departure] of night, [81:17]

وَاللَّيْلِ إِذْ أَدْبَرَ

And the night / when / it departs ,/

And the night when it withdraweth [74:33]

[1] See for example: http://quran.com and http://qurandatabase.org

[2] http://corpus.quran.com

[3] http://textminingthequran.com/wiki

[4] We give the Arabic text with English word-by-word translation available at http://corpus.quran.com followed by Pickthall translation available at http://quran.com

3. In the following pair of verses, the first indicates indefinite "words of revelation" which are explicitly mentioned in the second verse:

فَتَلَقَّىٰ آدَمُ مِن رَّبِّهِ كَلِمَاتٍ فَتَابَ عَلَيْهِ ۚ إِنَّهُ هُوَ التَّوَّابُ الرَّحِيمُ

Then received / Adam / from / his Lord / words, / So (his Lord) turned / towards him. / Indeed He! / He / (is) the Oft-returning (to mercy), / the Most Merciful. /

Then Adam received from his Lord words (of revelation), and He relented toward him..[2:37]

قَالَا رَبَّنَا ظَلَمْنَا أَنفُسَنَا وَإِن لَّمْ تَغْفِرْ لَنَا وَتَرْحَمْنَا لَنَكُونَنَّ مِنَ الْخَاسِرِينَ

Both of them / "Our Lord / we have wronged / ourselves, / and if / not / You forgive / [for] us / and have mercy (on) us, / surely, we will be / among / the losers/ ".

They said: Our Lord! We have wronged ourselves. If thou forgive us not and have not mercy on us, surely we are of the lost! [7:23]

4. Finally, in the following pair, the term "*Lord of the Worlds*" in explained in the second verse, as "*Lord of the heavens and the earth and all that is between*";

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ

All praises and thanks / (be) to Allah, / the Lord / of the universe/

Praise be to Allah, Lord of the Worlds,[1:2]

قَالَ فِرْعَوْنُ وَمَا رَبُّ الْعَالَمِينَ قَالَ رَبُّ السَّمَاوَاتِ وَالْأَرْضِ وَمَا بَيْنَهُمَا ۖ إِن كُنتُم مُّوقِنِينَ

Firaun said / Firaun said / "And what / (is the) Lord / (of) the worlds?" / He said, / "Lord / (of) the heavens / and the earth / and whatever / (is) between them, / if / you (should) be / convinced." /

Pharaoh said: And what is the Lord of the Worlds? (Moses) said: Lord of the heavens and the earth and all that is between them, if ye had but sure belief. [26:23,24]

It is clear from the above examples that relatedness in the Quran needs to be comprehended in a very broad sense, much broader than relatedness on just a lexical or word level. Often deep understanding and domain knowledge is necessary to relate two otherwise seemingly unrelated verses.

2. Compilation Process

2.1 Previous Datasets

There have been a number of data sources used previously for computational analysis of paraphrased texts and sentence similarity. The Multiple-Translation Chinese Corpus (Huang et at 2002) contains 11 English translations of 105 news stories in Mandarin Chinese, which amounts to 993 sentences. (Li et al. 2006) produced a dataset comprising 65 pairs of sentences with similarity scores from 32 human judges. Microsoft Research released a corpus of paraphrased text containing 5801 pairs of sentences collected from various news sources (Dolan et al. 2004). Among this set 3,900 pairs are considered "semantically equivalent" by human judges. Each pair has been visited by at least two annotators with an average agreement rate of 83%.

As for similarity and relatedness at word level where human judges provided scores we can cite: (Rubenstein & Goodenough, 1965) provided a dataset of 65 word pairs, (Miller & Charles, 1991) provided a dataset of 30 word-pairs, and (Finkelstein et al., 2002) with 353 word pairs. Experiments in word relatedness often use datasets from SAT test questions where the most related word pairs need to be selected from five candidates.

2.2 Ibn Kathir's Tafsir

Ismail Ibn Kathir was a Muslim scholar who died in 1373 CE, well known for his classic book of Quran commentary (or *Tafsir* in Arabic). This book is one of the most widely cited commentaries of the Quran. Ibn Kathir followed a regular methodology when commenting on a verse, which he made clear at the introduction of his book. Firstly, he discusses other related verses explaining the current verse. Often, when a certain verse covers a subject briefly, there might be many other verses that cover other aspects of this subject; see the examples in section 1.3 above. Secondly, he refers to traditions and saying of the Prophet Muhammad (i.e., *Hadith*). Thirdly, he cites opinions of *Sahabah* (i.e., companions of the Prophet) on this verse, especially those who are well known for their knowledge of the Quran like *Ibn Abbas* and *Ibn Masoud*.

We exploited this methodology for the purpose of creating a dataset of related verses. We understand that Ibn Kathir never claimed to exhaustively cite all related verses when commenting on a particular verse, nor did he always observe the commutative property of relatedness, i.e., if verse y was cited while commenting on verse x, then x should also appear at the commentary page of verse y. These observations allowed us to expand the original list of related verses beyond what is found in Ibn Kathir.

2.3 Dataset Compilation

Tafsir Ibn Kathir is available online at several websites. We chose the online version available at the official website of the King Fahd Complex for the Printing of the Holy Qur'an[5]. After observing the structured format used in this site for displaying this *Tafsir*, we developed scripts to extract verse pairs automatically. Our script automatically retrieved the URL of a given verse and extracted the chapter and verse numbers of all other verses mentioned in the context of a given verse.

After the initial compilation of the dataset, manual

[5] http://qurancomplex.com

intervention was necessary to clean up a few inconsistencies, as well as to adjust correct pairing of verses, because in the original Ibn Kathir's *Tafsir*, a group of verses are discussed at a time, while our dataset contains only verse pairs. Through this process we collected a total of 7,679 pairs of single verses which were then fed into relational database tables using MySQL.

2.3 Dataset filtration and extension

Upon initial investigation of the dataset, we realized that a second manual check was required to filter the dataset further for it to be useful for the intended computational analysis tasks. While commenting on a particular verse, Ibn Kathir's discussion might lead to a distant topic, making the task of computation of the relatedness almost impossible.

Consider for example, the following pair from our dataset, where no obvious relation is found before reading its context in Ibn Kathir:

إِنَّ اللَّهَ لَا يَسْتَحْيِي أَن يَضْرِبَ مَثَلًا مَّا بَعُوضَةً فَمَا فَوْقَهَا ۚ فَأَمَّا الَّذِينَ آمَنُوا فَيَعْلَمُونَ أَنَّهُ الْحَقُّ مِن رَّبِّهِمْ ۖ وَأَمَّا الَّذِينَ كَفَرُوا فَيَقُولُونَ مَاذَا أَرَادَ اللَّهُ بِهَٰذَا مَثَلًا

Indeed, / Allah / (is) not / ashamed / to / set forth / an example / (like) even / (of) a mosquito / and (even) something / above it. / Then as for / those who / believed, / [thus] they will know / that it / (is) the truth / from / their Lord. / And as for / those who / disbelieved / [thus] they will say / what / (did) intend / Allah / by this / example?

Lo! Allah disdaineth not to coin the similitude even of a gnat. Those who believe know that it is the truth from their Lord; but those who disbelieve say: What doth Allah wish (to teach) by such a similitude? [2:26]

فَلَمَّا نَسُوا مَا ذُكِّرُوا بِهِ فَتَحْنَا عَلَيْهِمْ أَبْوَابَ كُلِّ شَيْءٍ حَتَّىٰ إِذَا فَرِحُوا بِمَا أُوتُوا أَخَذْنَاهُم بَغْتَةً فَإِذَا هُم مُّبْلِسُونَ

So when / they forgot / what / they were reminded / of [it], / We opened / on them / gates / (of) every / thing, / until / when / they rejoiced / in what / they were given, / We seized them / suddenly / and then / they / (were) dumbfounded. /

Then, when they forgot that whereof they had been reminded, We opened unto them the gates of all things till, even as they were rejoicing in that which they were given, We seized them unawares, and lo! they were dumbfounded. [6:44]

Ibn Kathir drew an analogy between the situation of a gnat (or mosquito) who overfeeds itself till death, to those people who wrongly over-enjoy the provision of this world till God's punishment befalls them.

Considering these kinds of examples, the entire dataset was manually checked and - instead of completely removing these pairs - a special 'not obvious' flag was placed against 883 such cases. With the remaining 6,796 semantically related verse pairs, we believed further distinctions in the degree of relatedness were needed if the dataset were to be used for training learning algorithms. Consider for example verse 78:20 quoted above in section 1.3; Ibn Kathir cited the following three consecutive verses in his commentary on 78:20. While the first verse 20:105 is strongly related to 78:20, the other two complete the picture in the context.

وَيَسْأَلُونَكَ عَنِ الْجِبَالِ فَقُلْ يَنسِفُهَا رَبِّي نَسْفًا

And they ask you / about / the mountains, / so say, / "Will blast them / my Lord / (into) particles. /

They will ask thee of the mountains (on that day). Say: My Lord will break them into scattered dust. [20:105]

فَيَذَرُهَا قَاعًا صَفْصَفًا

Then He will leave it, / a level / plain. /

And leave it as an empty plain, [20:106]

لَّا تَرَىٰ فِيهَا عِوَجًا وَلَا أَمْتًا

Not / you will see / in it / any crookedness / and not / any curve." /

Wherein thou seest neither curve nor ruggedness. [20:107]

Thus, a second scrutiny of the dataset resulted in assigning two levels of degree of relatedness: level 2 (in total 3,079 pairs) represents strong relations as between verses 78:20 and 20:105, and level 1 (total 3,718 pairs) represents weaker relations as between 78:20 and 20:106 above. Manual filtration of all levels described above was performed by the first author.

We suggest two ways in which this dataset could be extended: (a) for a pair of strongly related verses <x,y> (i.e., level 2) the pair <y,x> should be included if not already in the dataset. (b) Consider a related pair <x,y>, if <y,z> is also strongly related, then both <x,z> and <z,x> could be added as well.

وَسُيِّرَتِ الْجِبَالُ فَكَانَتْ سَرَابًا

And are moved / the mountains / and become / a mirage. /

And the hills are set in motion and become as a mirage. [78:20]

وَيَوْمَ نُسَيِّرُ الْجِبَالَ

And the Day / We will cause (to) move / the mountains/

And (bethink you of) the Day when we remove the hills..[18:47]

وَتَسِيرُ الْجِبَالُ سَيْرًا

And will move away, / the mountains / (with an awful) movement /

And the mountains move away with (awful) movement [52:10]

As an illustration, consider the three verses above. We find that <78:20, 18:47> is a level 2 pair in our dataset. However, <18:47, 78:20> is not found but could be added as a new pair. Similarly, we notice that <18:47, 52:10> is a level 2 pair in the dataset, however, the pair <78:20, 52:10> was not considered by Ibn Kathir neither was the

pair <52:10,78:20>, and both could be added as strongly related verses.

The Quranic Arabic Corpus (QAC) gives the root of each word, so we used this to count the availability of matching lexical roots in all paired verses in our dataset.

The dataset was made available for download as XML file containing the following format[6] (see Appendix):

```
<column name="uid">1</column>
<column name="ss">1</column>
<column name="sv">1</column>
<column name="ts">1</column>
<column name="tv">2</column>
<column name="common">0</column>
<column name="relevance">2</column>
```

Denoting that a pair of verses `<ss:sv, ts:tv>` has `common` number of matching lexical roots and are related with a degree of `relevance`, which could be 0,1 or 2. The available file contains the original 7,679 pairs, while the extension of the dataset could be made computationally following the logic described above. We kept only reference to chapter and verse numbers since most electronic version of the original Quran text as well as its translations maintain these references.

3. Applications using QurSim

3.1 Online Visualization

The QurSim dataset has been captured as MySQL in order to enable web queries and visualization. We have created online query pages[7] where the user inputs a verse number and is returned with both directly and indirectly related verses, in Arabic and English, along with the degree of relatedness and common roots as shown in figure 1. Moreover, thanks to integration with QurAna - pronominal anaphora corpus - (Sharaf & Atwell, 2012), we provided information on concepts as antecedents of pronouns in each verse, as well as a concept cloud (see figure 2) from all verses, given at the end to give the user an idea of the major concepts involved.

Figure 3 shows how our online application enables better visualization of related verses using Dracula Graph Visualization tool[8]. Each node represents a verse, and the arrows show the number of common root words between the related verses. We believe these query tools designed using QurSim dataset will benefit Quranic students and researchers alike. With the ability to visualize direct and indirect links of a verse, researchers will be able to relate verses easily and supplement Ibn Kathir's initial list.

[6] Can be downloaded from: http://www.textminingthequran.com/wiki/Verse_relatedness_in_Ibn_Kathir

[7] http://www.textminingthequran.com/apps/similarity.php

[8] http://www.graphdracula.net

Following are 5 verses directly related to 7:187 from Ibn Kathir:

No.	Arabic	English	Common Roots	Level
33:63	يَسْأَلُكَ النَّاسُ عَنِ السَّاعَةِ قُلْ إِنَّمَا عِلْمُهَا عِندَ اللَّهِ وَمَا يُدْرِيكَ لَعَلَّ السَّاعَةَ تَكُونُ قَرِيبًا	Men ask thee of the Hour. Say: The knowledge of it is with Allah only. What can convey (the knowledge) unto thee? It may be that the Hour is nigh. **Pronoun Referents:** Prophet Muhammad, the Hour.	11	2
31:34	إِنَّ اللَّهَ عِندَهُ عِلْمُ السَّاعَةِ وَيُنَزِّلُ الْغَيْثَ وَيَعْلَمُ مَا فِي الْأَرْحَامِ وَمَا تَدْرِي نَفْسٌ مَّاذَا تَكْسِبُ غَدًا وَمَا تَدْرِي نَفْسٌ بِأَيِّ أَرْضٍ تَمُوتُ إِنَّ اللَّهَ عَلِيمٌ خَبِيرٌ	Lo! Allah! With Him is knowledge of the Hour. He sendeth down the rain, and knoweth that which is in the wombs. No soul knoweth what it will earn to-morrow, and no soul knoweth in what land it will die. Lo! Allah is Knower, Aware. **Pronoun Referents:** Allah.	7	2
42:18	يَسْتَعْجِلُ بِهَا الَّذِينَ لَا يُؤْمِنُونَ بِهَا وَالَّذِينَ آمَنُوا مُشْفِقُونَ مِنْهَا وَيَعْلَمُونَ أَنَّهَا الْحَقُّ أَلَا إِنَّ الَّذِينَ يُمَارُونَ فِي السَّاعَةِ لَفِي ضَلَالٍ بَعِيدٍ	Those who believe not therein seek to hasten it, while those who believe are fearful of it and know that it is the Truth. Are not they who dispute, in doubt concerning the Hour, far astray? **Pronoun Referents:** the Hour, (Kaafir) the infidels, the believers , those who deny resurrection.	4	2
79:42	يَسْأَلُونَكَ عَنِ السَّاعَةِ أَيَّانَ مُرْسَاهَا	They ask thee of the Hour: when will it come to port? **Pronoun Referents:** the Hour, Prophet Muhammad, the infidels of Quraish.	4	2
21:38	وَيَقُولُونَ مَتَىٰ هَٰذَا الْوَعْدُ إِن كُنتُمْ صَادِقِينَ	And they say: When will this promise (be fulfilled), if ye are truthful? **Pronoun Referents:** the infidels of Quraish, Prophet Muhammad and the believers.	2	2

Figure 1. Verses directly related to 7:187

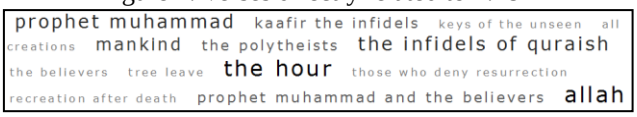

Figure 2: Concept cloud from pronoun referents of all related verses to 7:187

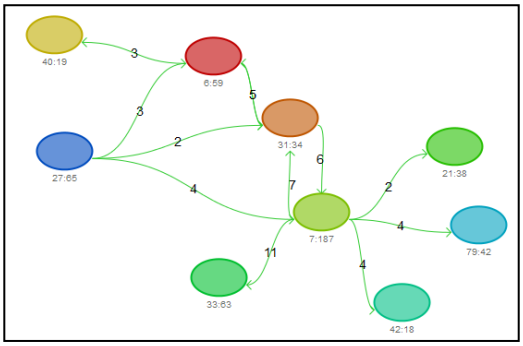

Figure 3: A network of verses related to 7:187

Domain experts can utilize this dataset for more interesting investigations in Quranic studies. For example, a Quranic student might want to find relatedness between Chapters rather than verses. Using QurSim, we can relate two chapters by the frequency of cross-reference between their verses. Figure 4 is an application that shows such relations. Nodes in this graph show chapters and the number over arrows show the number of cross-referenced verses between the chapters. Quranic chapters are broadly categorized thematically into Meccan or Medinan chapters distinguished in our graph as red or green respectively.

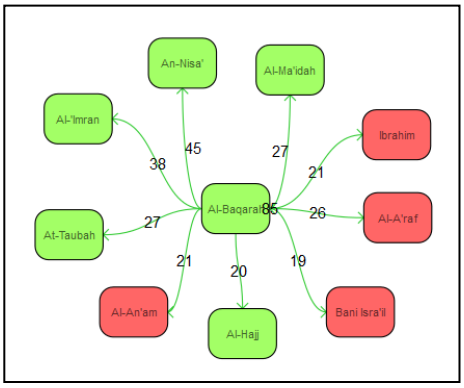

Figure 4: Relatives of chapter No. 2 "Al-Baqarah"

3.2 Verse Distance using Vector Space Model

The vector space model is widely used in information retrieval where the distance between the query terms and each document, represented as vectors, is measured by comparing the cosine of the angle between the vectors. We followed the same methodology and considered each verse of the Quran as a separate document.

Each verse then was modeled as term vectors taking roots of the Quran. The Quran has 1,226 unique roots, from these we have kept roots repeated over 2 times, and removed the first 3 most frequent roots. Thus, our vector for each verse contains 758 roots as term indices. Next, in order to assign a weight for each term, we used the term frequency – inverse document frequency (*tf-idf)* approach, using the following formula adapted from (Sebastiani 2002):

$$tfidf(t_k, d_j) = \#(t_k, d_j) \cdot \log \frac{|Tr|}{\#_{Tr}(t_k)}$$

Where$\#(t_k,d_j)$ denotes the number of times the root *tk* occurs in the verse d_j, and $\# T_r(t_k)$ denotes the verse frequency of root t_k, that is, the number of verses in the Quran T_r in which the root t_k occurs.

In order for the weights to fall in [0,1] interval and for the verses to be represented by vectors of equal length, the weights (w_{kj}) resulting from *tfidf* were normalized according to the following formula for *cosine normalization*:

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(tfidf(t_s, d_j))^2}}$$

To find the distance (or measure of similarity) between two vectors, cosine angle is measured using the formula below, where A, B denote two verses' vectors:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

Values are between [0,1], where 0 indicates no similarity, and 1 indicates identical matching. Using the above setup, we have evaluated the 7,679 verse pairs in QurSim and found out that only 428 pairs (6%) produced similarity value above 0.5. This finding confirms the assumption that automatic computation of verse relatedness requires integration with domain specific knowledge sources, and relying on lexical matching only often produces poor results .

Next, to improve this low recall, we considered how to enrich a verse vector with pronoun referents which we developed earlier (Sharaf and Atwell 2012). Instead of constructing vectors for a verse from only that verse's roots, we augmented this verse's roots with roots of all other verses that share common antecedent. For example consider verse 27:26 below:

بَلِ ادَّارَكَ عِلْمُهُمْ فِي الْآخِرَةِ ۚ بَلْ هُمْ فِي شَكٍّ مِّنْهَا ۖ بَلْ هُم مِّنْهَا عَمُونَ

Nay, / is arrested / their knowledge / of / the Hereafter? / Nay / they / (are) in / doubt / about it. / Nay, / they / about it / (are) blind. /

Nay, but doth their knowledge reach to the Hereafter? Nay, for they are in doubt concerning it. Nay, for they cannot see it.

This verse contains 3 concepts marked by pronoun referents: 'the polytheists', 'those who deny resurrection' and 'the world Hereafter'. Therefore, we have augmented the term vector of the verse 27:66 with the terms from all other verses that have any of these three concepts.

The similarity measurement experiment described above was repeated using these improved vectors, and the same dataset was used. While in the early experiment, only 428 pairs showed similarity distance over 0.5, augmenting verses with their concepts showed 869 pairs from the total of 7,679 pairs in our dataset had similarity over 0.5, i.e., over 50% improvement.

4. Challenges

As Quranic verses vary in size we run into two different problems: 1) Those verses that are long may cover several topics and hence, pairing the whole verse with another verse reflects only a partial relation 2) those verses that are very small share with adjacent verses a single topic, and again in this case the one-to-one pairing with another verse is not appropriate. As an example for the first point, consider verses related to verse 2:255 below. This is a relatively long verse containing 10 short sentences covering different aspect of Allah's attributes and quality. Ibn Kathir links this verse with 15 different verses.

اللَّهُ لَا إِلَٰهَ إِلَّا هُوَ الْحَيُّ الْقَيُّومُ ۚ لَا تَأْخُذُهُ سِنَةٌ وَلَا نَوْمٌ ۚ لَّهُ مَا فِي السَّمَاوَاتِ وَمَا فِي الْأَرْضِ ۗ مَن ذَا الَّذِي يَشْفَعُ عِندَهُ إِلَّا بِإِذْنِهِ ۚ يَعْلَمُ مَا بَيْنَ أَيْدِيهِمْ وَمَا خَلْفَهُمْ ۖ وَلَا يُحِيطُونَ بِشَيْءٍ مِّنْ عِلْمِهِ إِلَّا بِمَا شَاءَ ۚ وَسِعَ كُرْسِيُّهُ السَّمَاوَاتِ وَالْأَرْضَ ۖ وَلَا يَئُودُهُ حِفْظُهُمَا ۚ وَهُوَ الْعَلِيُّ الْعَظِيمُ

Allah - / (there is) no / God / except / Him, / the Ever-Living, / the Sustainer of all that exists. / Not / overtakes Him / slumber / [and] not / sleep. / To Him (belongs) / what(ever) / (is) in / the heavens / and what(ever) / (is) in / the earth. / Who / (is) the one / who / can intercede / with Him / except / by / He knows / what / (is) / before them / and what / (is) behind them. / And not / they encompass / anything / of / His Knowledge / except / [of] what / He willed. / Extends / His Throne / (to) the heavens / and the earth. / And not / tires Him / (the) guarding of both of them. / And He / (is) the Most High, / the Most Great. /

Allah! There is no deity save Him, the Alive, the Eternal. Neither slumber nor sleep overtaketh Him. Unto Him belongeth whatsoever is in the heavens and whatsoever is

in the earth. Who is he that intercedeth with Him save by His leave? He knoweth that which is in front of them and that which is behind them, while they encompass nothing of His knowledge save what He will. His throne includeth the heavens and the earth, and He is never weary of preserving them. He is the Sublime, the Tremendous.

As an example on the second point, consider verse 11:97 below. Ibn Kathir referred to six consecutive small verses as related to this verse. Since we paired one single verse with another, in our dataset this relation is represented by six pairs <11:97,79:21>, <11:97, 79:22>, … <11:97,79:26>. Note how the last pair <11:97, 79:26> is very weakly related when taken in isolation.

إِلَىٰ فِرْعَوْنَ وَمَلَئِهِ فَاتَّبَعُوا أَمْرَ فِرْعَوْنَ ۖ وَمَا أَمْرُ فِرْعَوْنَ بِرَشِيدٍ

To / Firaun / and his chiefs, / but they followed / (the) command of Firaun, / and not / (the) command of Firaun / / was right. /

Unto Pharaoh and his chiefs, but they did follow the command of Pharaoh, and the command of Pharaoh was no right guide.[11:97]

فَكَذَّبَ وَعَصَىٰ ثُمَّ أَدْبَرَ يَسْعَىٰ فَحَشَرَ فَنَادَىٰ فَقَالَ أَنَا رَبُّكُمُ الْأَعْلَىٰ فَأَخَذَهُ اللَّهُ نَكَالَ الْآخِرَةِ وَالْأُولَىٰ إِنَّ فِي ذَٰلِكَ لَعِبْرَةً لِّمَن يَخْشَىٰ

But he denied / and disobeyed. / Then / he turned his back, / striving, / And he gathered / and called out, / Then he said, / "I am / your Lord, / the Most High." / So seized him / Allah / (with) an exemplary punishment / (for) the last / and the first. / Indeed, / in / that / surely (is) a lesson / for whoever / fears. /

[21]But he denied and disobeyed, [22]Then turned he away in haste, [23]Then gathered he and summoned, [24]And proclaimed: " I (Pharaoh) am your Lord the Highest." [25] So Allah seized him (and made him) an example for the after (life) and for the former. [26] Lo! herein is indeed a lesson for him who feareth. [79: 21-26]

Another challenge we face is when Ibn Kathir elaborates on a particular word from a verse and brings in different verses in the course of explanation. These cited verses might not seem related without relating back to the context made in Ibn Kathir. For example consider the verse 11:8 where the word "Ummah" was mentioned, which means a "nation". However, in the Quran this word can have other less frequently used meanings like "a leader" or "a short period of time". Here Ibn Kathir cites references of all other verses in the Quran where this word is used to mean things other than a "nation".

5. Conclusion

We have presented QurSim as a language resource for Quranic scholars, students and for researchers in the computational linguistics community, particularly those investigating computational text relatedness measures. QurSim was built relying on scholarly works to guarantee the quality of the data. The dataset could be improved further. As Quranic verses vary in size, a pair of two large size verses might relate based on a small phrase within these verses. Such instance of pairs could be cropped so only related phrases are preserved. Books of Tafsir other than Ibn Kathir could be consulted to increase the size of our dataset. Traditions of Prophet Muhammad narrated to explain verses could also be incorporated from Ibn Kathir to enrich this dataset.

Computational analysis of text relatedness is a growing research area. The lack of proper evaluation datasets stands as a major obstacle for progress in this field. Because of the availability of machine readable Quran translations in multiple languages, QurSim can potentially contribute in producing quality datasets in multiple languages and with minimum effort.

6. References

Dolan, W.; Quirk, C.; and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. *In Proceedings of the 20th International Conference on Computational Linguistics*.

Dukes, K.; Atwell, E.; and Sharaf, A. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Dependency Treebank. *Proc LREC'2010*, Valetta, Malta.

Finkelstein, L., Gabrilovich, E.,Matias, Y., Rivlin, E., Solan, Z.,Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1), 116–131

Gabrilovich, E.; and Markovitch. S., (2007). Computing semantic related-ness using wikipedia-based explicit semantic analysis. *In Proceedings of the 20th International Joint Conference on Arti¯cial Intelligence*, January.

Huang, S.; Graff, D.; and Doddington, G. (2002) Multiple-Translation Chinese Corpus. *Linguistic Data Consortium*, Philadelphia

Li, Y.; McLean, D.; Bandar, Z.A.; O'Shea, J.D.; Crockett, K.; (2006) Sentence Similarity Based on Semantic Nets and Corpus Statistics. *IEEE Transactions on Knowledge and Data Engineering*. vol. 18, no. 8, pp. 1138-1150

Miller, G., & Charles, W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

Resnik, P.; Olsen, M.B.; and Diab, M. (1999) The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues, *Computers and the Humanities*, 33(1-2), pp. 129-153.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comp. Survey*, 34(1):1-47.

Sharaf, A., Atwell, E. (2012) QurAna: Corpus of the Quran annotated with Pronominal Anaphora. *LREC 2012*, Istanbul, Turkey.

Appendix

Start of download file http://www.textminingthequran.com/wiki/File:Kathir-verse-similarity.xml

```
<!--
* PLEASE DO NOT REMOVE OR CHANGE THIS COPYRIGHT BLOCK
*======================================================================
*  Dataset on Quranic Verse Relatedness from Tafsir Ibn Kathir (version 0.1)
*  Copyright (C) 2011 Abdul-Baquee M. Sharaf
*  License: GNU Public License
*
*  This dataset lists pairs of verses that have been identified by Ibn
*  Kathir in his Tafsir book. After collecting these pairs, two further
*  passes were made manually to brand degree of relatedness.
*
*  Level '0':
*  seems very loosely related and should be understood by looking
*  into the context in the tafsir book.
*
*   Level '1':
*   These pairs are understandable by Human reader to be related, but
*   still might be difficult for training learning algorithms
*
*   Level '2':
*   These pairs are very much related and might be suitable for taining
*   machine learning algorithms.
*
*  TERMS OF USE:
*
*  - Permission is granted to copy and distribute verbatim copies
*    of this file, but CHANGING IT IS NOT ALLOWED.
*
*  - This annotation can be used in any website or application,
*    provided its source (TextMiningTheQuran.com) is clearly
*    indicated.
*
*  - This copyright notice shall be included in all verbatim copies
*    of the text, and shall be reproduced appropriately in all works
*    derived from or containing substantial portion of this file.
*
*  Check updates at (http://TextMiningtheQuran.com)
*
*  USAGE:
*
*   "uid"    : incremantal ID
*   "ss"     : source chapter number
*   "sv"     : source verse number
*   "ts"     : target chapter number
*   "tv"     : target verse number
*   "common" : the number of common root words between the two verses
*   "relevance"   : the degree of relatedness as explained above
-->
<pma_xml_export version="1.0">
    <database name="related-verses">
        <!-- Table kathir -->
        <table name="kathir">
            <column name="uid">1</column>
            <column name="ss">1</column>
            <column name="sv">1</column>
            <column name="ts">1</column>
            <column name="tv">2</column>
            <column name="common">0</column>
            <column name="relevance">2</column>
        </table>
```