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