Abs-270

Abdul-Baquee M. Sharaf and Eric Atwell

Creating a gold standard corpus for related texts

Text similarity and text relatedness are very important applications of corpus linguistics [1][2]. With the surge of textual information over the web and elsewhere, it is very hard to link similar or related information. As a result most search engines will not be able to find relations between related texts where lexical matching is absent.

In order to enable machine learning techniques to automatically detect similar and related texts, it is important first to create a corpus of training data that collects a large sample of similar and related texts which has been dictated as such by expert human sources. This would then form a gold standard for researchers to benchmark their machine learning algorithms.

In this paper we report our work on preparing a multilingual corpus of related texts. As a source we chose the Quran for several reasons. First, it has been a very popular text and considered sacred words of God by more than 1.5 billion Muslims worldwide. Second, over centuries there has been hundreds of scholarly works on the Quran ranging between explanatory commentaries to critiques to various translations in and within multiple languages. Third, the Quran is structured as chapters which in turn are broken into verses of short texts. As most information available today in the web or elsewhere are short text snippets, gathering related Quranic verses gives a reasonable good platform for researchers in the field of text relatedness. Fourth, the Quran is characterized as being highly cohesive text where related concepts within verses are repeated in various chapters. This feature allows populating a large sample of short text data for the purpose of text similarity studies.

There has been a large number of scholarly commentaries on the Quran know as Quranic exegesis or books of Tafsir. One popular such books is Tafsir Ibn Kathir authored by Ibn Kathir around 1350 AD [3]. This volume is characterized by pointing out similar and related verses in the context of commenting on a particular verse. Our work at this stage depended on Ibn Kathir in collecting related verses and annotation them with their reference numbers. Then a second pass ran through the collected data and expanded the list of related verses by including "distant relations" as well. If verse x is directly related to verse y, and verse y is directly related to verse z (according to Ibn Kathir), then we consider verse z a distant relative to verse x. Finally, we could easily expand this corpus to multilingual since the availability of Quran translations in hundreds of other languages [4]. We disseminated the results online which allows users to find all related verses of a chosen verse. Our current collection gives over 8,000 relations between the 6,236 total verses of the Quran. Below is a high level architectural depiction of our corpus.

FIGURE HERE

To the best of our knowledge our work is first of its kind. We anticipate that our data can serve a group of researchers. It will help Quranic researchers to search for information buried in related verses. It will also serve NLP researchers to train their machine learning algorithms for detecting text relatedness in multiple languages. Moreover, this data will be a valuable resource for researchers in translation studies to measure the quality of translation and observe how two related texts differ in different languages or within the same language. In future we look forward to include related verses from other books of Tafsir. We also want to expand our corpus with related verses from the Bible.

References

[1] Gabrilovich, E. and S. Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. Proceedings of IJCAI, 1606-1611.

[2] Mihalcea,R., C. Corley, and C. Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of AAAI-06.

[3] Ibn-Katheer (2006). Tafseer Al-Quran (In Arabic). Dar Al-Kutub al-Elmiyyah.

[4] The website http://qurandatabase.og for example gives over 100 translations in 55 languages