Similarity Measures for Classical Arabic Poetry Ranking

Iqbal A. Mohammed

College of Technical Medical & Health, Foundation of Technical Education, Ministry of Higher Education & Scientific Research, Baghdad, Iraq

iqbalabaki@gmail.com

Abstract. Arabic language is a very complicated language. The Classical Arabic poetry has difficult words in their meanings and grammar. Although these poetries use a very complex Arabic morphology it is not a tedious task to classify them. This approach is presenting a method to rank the poetries of a certain age in the Arabs history to their main ranks (like: Ritha'a, Hija'a, Ghazal…etc,) using the algorithm of N-grams statistical technique. The whole work is depending on normalizing the poetries without using any stemming process and using the n-gram (bi-grams) for retrieving characters. Retrieval influence tests have been implemented using similarity coefficients: the Dice, the Overlap and the Jaccard coefficient measures. For system evaluation Precision, Recall and F-measure have been calculated. Despite the fact that the Overlap measure seemed to be the best measure the rest gave good results as well.

Keywords: Information retrieval, N-gram, similarity measures, F-measure, ranking, normalization.

1 INTRODUCTION

The Classical Arabic Poetry (CAP) is written in a certain way. A verse is consisted of two sides in one line this is called (bayt). Most of the CAP using very complicated and different Arabic words, however there is no intend to deal with the language morphology and its complexity. Text ranking is the process of structuring a set of poems according to a group structure that is known in advance. In order to implement this approach the technique of n-gram statistical frequency is used. It provides a simple and reliable way in implementation. It can be found if two subsets are semantically similar or dissimilar from the structures of characters of these words.

The idea of this approach is rather confusing in collecting the related works to this work, as most of the works are using stemming process (to omit suffixes and prefixes from a word and reach the word root) before applying the n-gram technique. This work has overcome this step trying to discover what might happen and explore all possible ways that lead to better implementation. Not only this, it is rather challenging when using the bigrams instead of trigrams for the Arabic language mainly. This language most of its word roots are three letters roots and when using the trigrams, it will be quite adequate to reach to the word root without using stemming process. All related works reported that trigrams gave very good results whereas this work is concentrating on bigrams only.

Measuring similarity or distance between two entities is a key step for several data mining and knowledge discovery tasks. This work used similarity measures as many works used them ignoring the idea of using the dissimilarity ones as the last reported poor results with accordance to similarity measures.

The advantage of this approach is to investigate the characteristics of n-gram on CAP and to evaluate in terms of retrieval effectiveness of bi-gram ranking technique.

This approach is organized as follows: in section two the normalization process will be discussed, section three the N-gram technique is presented. In section four the similarity measures are briefly clarified. The whole approach is explained in section five. Finally the results followed by discussion and conclusion are discussed.

2 ARABIC NORMALIZATION TECHNIQUE

Arabic orthography is highly variable. For instance, changing the letter ي to ى at the end of the word is very common. Also changing آ,أ,إ are written as ا this is because of their similarity in appearance.

Arabic language have got 28 letters and written from right to left. Many difficulties appears when dealing with Arabic language like various spelling of certain words, irregular and inflected derived forms, short diacritics and long vowels, and most of its words have got affixes.

When a normalization process applied to Arabic text it has to make the text file compact. This process omit all the punctuation and diacritics like (ً, ْ, ٍ, ِ, ٌ, َ, ُ) from a text also the tatweel character (—), stop words which include Arabic prefixes, pronouns, and prepositions like (على, كيف, نحن, أين, الى, من), and foreign letters and numbers. But the most important letter to be removed is the conjunction و between any two words. Finally the following replacements are taking place in the normalization process:-

Replacing آ إ أ by alif bar ا.

Replacing ى by ي at the end of the words.

Replacing ة by ه at the end of the words.

Replacing the sequence of يء by ئ.

3 N-GRAM STATISTICAL TECHNIQUE

N-grams have been a popular area of research for many years and are most commonly used in empirical linguistics to build statistical models of language.

This method can be used for any language as it does not demand information about the language, also no limitation rules to be applied to the n-grams and there is no need to construct a vocabulary datasets. The meanings of the words have got no affect on this method.

The words are broken into character n-grams, in the case of not using a stemmer and the words are not stemmed, a match will be found between the training characters and the tested ones.

N-grams are created by applying a shifting window of n characters over a word. If the word has fewer than n characters, the whole word is returned.

The main benefits of the n-gram models that are used to solve many information retrieval problems are:

1- Language independence and simplicity: character level n-gram models are applicable to any language and even none language sequences such as music or gene sequences.
2- Robustness: n-grams are insensitive to spelling variations and errors, mainly in comparison to word feature.
3- Completeness: the vocabulary of character tokens is much smaller than any word vocabulary and normally is known in advance. Therefore the sparse data problem is much less serious in character n-gram models of the same order.

This approach suggested that words which have a high degree of structural similarity tend to be similar in meaning. Each word is represented by a list of its constituent n-grams, where n is the number of adjacent characters in the substrings. Typical values for n are 2 and 3 which correspond to the use of bigrams and trigrams. For bigram the number of n is n+1 and trigram is n+2. A quantitative similarity measure K between them can be computed by using Dice Coefficient or the Overlap Coefficient or Jaccard Coefficient as shown in table 1.

Table 1. Similarity measures of Bigram and Trigram for words سنديان, بنيان.

	Bigram	**Trigram**
Unique N-gram of word 1	_س سن ند دي يا ان ن_	__س _سن سند ندي ديا يان ان_ ن__
Unique N-gram of word 2	_ب بن ني يا ان ن_	__ب _بن نيا يان ان_ ن__
A= unique N-gram of word 1	7	8
B= unique N-gram of word 2	6	6
C= Shared unique N-gram	3	3
Dice coefficient= (2C)/(A+B)	0.4615385	0.4285714
Overlap coefficient = C/min(A,B)	0.5	0.5
Jaccard coefficient = C/|A|U|B|	0.3	0.27272727

4 DISTANCE AND SIMILARITY MEASURES

A similarity measure is a function which computes the degree of similarity between a pair of text object. It is used for calculating the distance between known profiles and the profile of the file to be ranked. For that, several measures of similarity can be used. This is done for the purpose of studying the influence of a system of ranking.

4-1 Jaccard Similarity Coefficient

It is used for comparing the similarity and diversity of sample sets.

The Jaccard coefficient measures similarity between sample sets, is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \, . \qquad (1)$$

It ranges between 0 and 1. It is 1 when A=B and 0 when A and B are disjoint, where 1 means the two sets are the same and 0 means they are completely different.

4-2 Overlap Similarity Coefficient

The overlap coefficient is a similarity measure that computes the overlap between two sets which is defined as follows:-

$$(A,B)=\frac{|A\cap B|}{\min(|A|,|B|)} . \qquad (2)$$

If set A is a subset of B or the converse then the overlap coefficient is equal to one.

4-3 Dice Similarity Measure

It is a similarity measure over sets. When taken as a string similarity measure, the coefficient may be calculated for two strings, A and B using bigrams as follows:-

$$S = \frac{2Nt}{Na + Nb} , \qquad (3)$$

Where Nt is the number of character bigrams found in both strings. Na is the number of bigrams in string A. Nb is the number of bigrams in string B.

5 EXPERIMENTAL DETAILS

It is obvious that any ranking system requires three kinds of processes; the first one is collecting the data and applies the normalization process to it, the second deriving the bigrams from the data collected, finally the third is to find out all ranked poems.

5-1 CAP Pre-Processing Phase

The advantage of this phase is to transform the CAP text file into a more suitable form and this for the purpose of ranking. Before implementing this phase the data have been collected from Arabic literature books. The poems have been collected so that they cover all the subsets they should belong to and for each subset a reference has been given to it.

For each dataset six poetries have been collected each have ten verses. This means that sixty verses for each subset. Only one text file has got all these poems for every subset. The several spaces between the two halves of a verse have been replaced by only one space, also there is no use for the enter key so that the text file would look like of one written line. This file would appear rather small.

The normalization process then, have been applied to the text file. The main advantage of using the CAP instead of Arabic corpora is that any corpus might have foreign letters, numbers and it is filled with punctuation. The normalization process (described in section two) makes the text file more compact.

5-2 Bigram Generation

The sixty verses that were collected in one text file have been separated as forty verses in one text file to act as the training set and the rest twenty in another text file taking the role of the testing set. The n-gram profile has been generated according to the work of Canvar.

The N-gram was computed (n=2 bigram) for all subset files. The spaces in the beginning and at the end of a bigram have been omitted, then all bigrams were inserted into a table to find the counter for the bigrams, it is incremented whenever a similar bigram occurred. The total number of occurrences has been saved in a different file and a descending sort for them has been applied. The final result is the profile of the text file for the training set and the testing set.

5-3 CAP Ranking Phase

In this phase the distance is measured. It is used to show how far out of place an N-gram in the training profile is from its place in the testing profile in terms of similarity. First thing to be done is calculating the rank order statistics for two profiles by measuring the difference in the positions of an N-gram in two different profiles. For each N-gram in the testing profile, a search for the same N-gram in the training rank profile has been done to calculate the difference between their positions. In case they are not found in the training rank profile, a maximum value is assigned. To end this process the sum of the distance measures for all N-grams are calculated. Using the three equations mentioned in the previous section the Jaccard, Overlap and the Dice similarity measures have been calculated. After computing the overall similarity measure between the testing profiles and with all the ranks training profiles, the rank that has the largest Dice, Jaccard and Overlap measure, is chosen as the rank that the testing profile belongs to.

The following example calculates the similarity between the words قارب , مقارب the set of bigrams in each word are { قا , ار , رب } { مق , قا , ار , رب } the intersection of these two sets has three elements for calculating the Dice measure it is found D= (2.3)/(3+4)= 0.857 and the Jaccard measure J=3/(3+4)-3=0.75 finally the Overlap measure O=3/3=1. Having this the final step in the experimental work is completed.

6 EVALUATION MEASURES AND RESULTS

For the purpose of evaluating the CAP ranking the recall and precision have been calculated for the three similarity measures. Precision is called a measure of exactness or fidelity. It is defined as a fraction of correctly categorized test cases divided by the number of test cases claimed to be similar. Precision= retrieved and relevant ranks/total retrieved ranks

$$P(A,B)=|A\cap B|/|A|. \qquad (4)$$

While the recall is a measure of completeness and it is defined as a fraction of correctly categorized test cases divided by the number of test cases manually categorized as similar. Recall= retrieved relevant ranks/total relevant ranks

$$R(A,B)=|A\cap B|/|B|, \qquad (5)$$

Where A is the retrieved set and B is the relevant set.

Also the F-measure which is the harmonic mean of precision and recall is calculated. It is a measure of a test's accuracy, 1 is its best value and 0 for its worst.

$$F(Recall,Precision)=2 * Recall * Precision/Recall + Precision. \qquad (6)$$

Tables 2,3,4 shows the computed values of precision, recall and F-measure for the three measures.

Table 2. Precision, Recall & F-measure using Jaccard's measure.

Ranks	Precision	Recall	F
Rank1"Ghazal"	0.727272	0.88888	0.79999
Rank2"Retha'a"	0.66666	0.11764	0.19998
Rank3"Madeh"	0.66666	0.28571	0.39999
Rank4"Heja'a"	0.5	0.33333	0.39999
Rank5"Wasef"	0.125	0.08333	0.09999
Rank6"Fakar"	0.05882	0.33333	0.1
Rank7"Nasseb"	0.15384	0.28571	0.19997
Rank8"Hekmah"	0.153846	0.28571	0.19999

Table 3. Precision, Recall & F-measure using Dice's measure.

Ranks	Precision	Recall	F
Rank1"Ghazal"	0.54545	0.66666	0.59999
Rank2"Retha'a"	0.18181	0.22222	0.19998

Rank3"Madeh"	0.7	0.7	0.7
Rank4"Heja'a"	0.30769	0.57142	0.39999
Rank5"Wasef"	0.14285	0.07692	0.09997
Rank6"Fakar"	0.125	0.5	0.2
Rank7"Nasseb"	0.33333	0.142857	0.19999
Rank8"Hekmah"	0.25	0.375	0.3

Table 4. Precision, Recall & F-measure using Overlap's measure.

Ranks	Precision	Recall	F
Rank1"Ghazal"	0.46153	0.85714	0.59999
Rank2"Retha'a"	0.1875	0.75	0.3
Rank3"Madeh"	0.35714	0.83333	0.49999
Rank4"Heja'a"	0.5	0.5	0.5
Rank5"Wasef"	0.4	0.13333	0.19999
Rank6"Fakar"	0.2	0.6	0.3
Rank7"Nasseb"	0.11111	0.09090	0.09994
Rank8"Hekmah"	0.28571	0.15384	0.2

7 DISCUSSION

It is obvious that the fourty verses that have been used as training data for each rank are very small. Although they were rich, well-balanced and hold deeply meaning features for each rank, this small size affects the process of ranking. Some of the ranks need more different verses in order to briefly cover all the words that might be needed to be with the bigram profile. The training data should be larger and cover wide range of each rank.

The bigram have been chosen to implement this work. This led to a different size of the profiles for both the training and the testing sets. It can be seen that some verses have very few words and others have many, with this the size of the text file that contains the data does not affect this work and it has got no influence on the whole work.

Some of the ranks retrieved many poems this means they have a high recall others retrieved very few non-relevant poems, which lead to having a very high precision. With this it can be noticed that this information retrieval system is a good one.

This approach used only similarity measures for their effectiveness instead of dissimilarity because many previous works proved that the results with similarity measures are better than the dissimilarity ones. Here we used the Jaccard, Overlap and Dice measures the three measures were effective and gave good results. Although the Overlap gave better results (for main ranks) on the whole, still the rest did their job and were effective in retrieving poems.

The F-measure chart shown in figure 1 below reflects the accuracy of the similarity measures in retrieving the poems of the eight ranks used in this system. It shows ranks like Ghazal, Madeh and Hija'a scored best values. As long as these values close to one, then the system is a very good one in retrieving poems. Some of the ranks showed a very low F value because of the difficulties in covering all the words that specify the rank which the testing poems belong to.

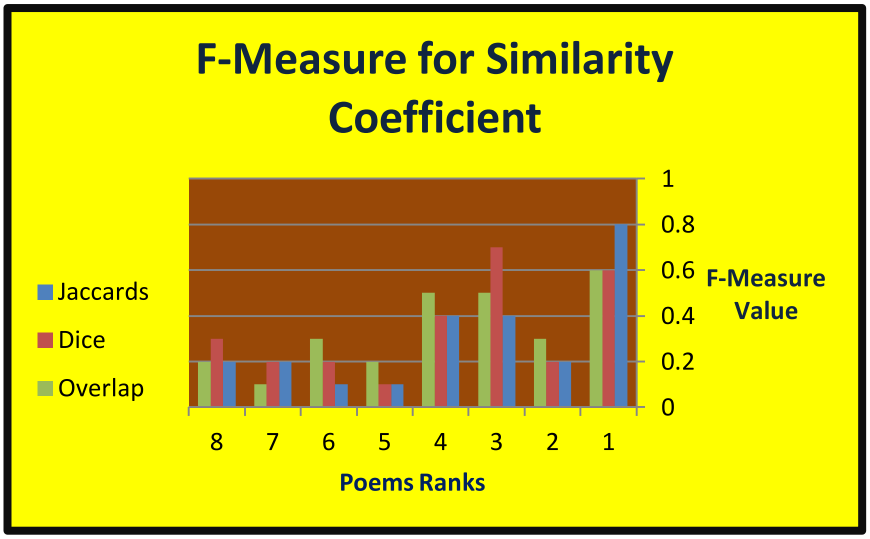

Fig. 1. F-Measure for Similarity Coefficients.

8 CONCLUSION

This paper presented a method for ranking Classical Arabic Poetry. The whole work depends on normalizing the verses of the poems and collecting them in one text file then generating the bigrams. Three measures of similarity have been used to show which one is the best in retrieving these poems. The Overlap measure showed better results although the Dice and the Jaccard worked properly. For future work it is recommended to use all the n-grams and make a comparison between them to show which gram is the best one.

Acknowledgments

Many thanks for Mr. Ayed (Lecturer) head of community health department, and Miss Jian, both at the college of technical & medical health for their help. My great appreciations for Dr. Shawqi Naji Jawad in Amman for encouraging me to work on this research.

REFERENCS

Khreisat, L. (2006). *Arabic Text Classification Using N-gram frequency Statistics a Comparative study*. The International Conference on data Mining Part of the World congress in Computer science DMIN: 78-82.

Al-Hajjar A., Hajjar M. and Zreik K. (2009). *Classification of Arabic Information Extraction methods*. Conference ICHSL7, Toulouse, pp. 311-317.

K.zreik, M.rammal, M. Sanan (2006). *Arabic Documents Classification using N-gram*. pp. 23-30.

K.zreik, M.rammal, M. Sanan "2008" *Arabic Supervised Learning Method Using N-gram.*

Chen, A. and gey, F. (2002). *Building an Arabic Stemmer for Information Retrieval*. TREC-11 Conference, pp 86-91.

I. Mohammad (2010). *Classical Arabic Poetry Categorization Using N-Gram Frequency Statistics*. Iraqi Journal of Science, Vol.51, no.1, pp.159-165.

T.M.T. Sembok, Z. AbuBakar (2011). *Effectiveness of Stemming and n-grams String Similarity Matching on Malay Documents*. International Journal of Applied Mathematics and Informatics, Issue 3, Volume 5, pp 208-215.

http://en.wikipedia.org/wiki/jaccard_index

A. Huang (April 2008). *Similarity Measures for text Document clustering.* Proceedings of the New Zealand Computer Science Research Student Conference. pp 49-56.

http:// en.wikipedia.org/wiki/Dice's_coefficient

جرجي زيدان, (1957). *تاريخ آداب اللغة العربية*, دار الهلال, ج1,ص 98,ج2 ص 105, ج3 ص33, ج4 ص 19, 105.

د. عمر فروخ، (1974). *المنهاج في الأدب العربي*, المكتبة العصرية, ج1, ص66, ج2, ص213, 216,219, ج3, ص32,33, ج4, ص 39, 37, 34.

د. هدى شوكت بهنام, *مقدمة القصيدة العربية في الشعر الأندلسي*, دار الشؤون الثقافية, ص 311, 325.

Canvar, W. and Trenkle, J. (1994). *N-gram-Based Text categorization*, in proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and information retrieval, pp 250-263.

Iqbal A. Mohammed is a lecturer at the Foundation of Technical Education. She received her BS and MS degrees in computer science from the University of Baghdad in 2001 and 2004, respectively. She is the author of three journal papers. Her current research interests include text mining, information retrieval, and NLP.