SINAI at MLIA COVID-19

José-Alberto Mesa-Murgado[1], Pilar López-Úbeda[1], and Manuel-Carlos Díaz-Galiano[1] María-Teresa Martin-Valdivia[1]

SINAI Research Group - CEATIC - Universidad de Jaén
Campus Las Lagunillas s/n. E-23071, Jaén, Spain
{jmurgado,plubeda,mcdiaz,maite}@ujaen.es

Abstract. This study describes the participation of the research group SINAI at the Universidad de Jaén, Spain. We have focused our efforts in solving the task 2 of the MLIA COVID-19 Workshop aimed at developing a Semantic Search System given a COVID-19 related corpus. For this purpose we submit 5 runs that use different features from the corpus in order to get different results.

1 Introduction

The outbreak of the Coronavirus disease 2019 (COVID-19) has led to a rapid and proactive response from medical and Artificial Intelligence (AI) communities worldwide. Research focused on Information Retrieval (IR) and Natural Language Processing (NLP) have concentrated their efforts on building datasets [7] and tools for efficiently managing the growing literature on COVID-19 [6] and other related diseases [3].

Inspired by the idea of IR and NLP research on COVID-19, we have participated in the MLIA COVID-19 Workshop Task2 which encourages the participants to develop a semantic search system that could be either monolingual or bilingual based. For the development of this system we have specifically focused on Spanish language. For this purpose, we have been provided with a COVID-19 related corpus that contains more than 800,000 documents written in Spanish.

The paper consists of 5 section that will be upgraded with more information after the end of every each one of the 3 rounds this workshop consists of. Therefore, Section 2 introduces the methodology that has been followed in order to the development of our semantic search system. Section 3 describes the experiments that we have run to test it as well as the features that involves each one of them. Section 4 discusses the results obtained in the evaluation of our system runs. To conclude, Section 5 brings a close to our participation as well as proposals for future work.

Covid-19 MLIA @ Eval Initiative, http://eval.covid19-mlia.eu/.

2 Methodology

2.1 Document format Identification

We have decided to choose Spanish as the only language for our information retrieval system but first, we have to identify the content of the attached XML documents files that build up the corpus.

Concerning the topics provided, we have 3 main aspects from which to classify the information received:

1. Based on the keywords associated to the document.
2. Based on the title of the document.
3. Based on the textual content included on the document.

This information is easy to extract from the XML documents which we have received and, although we could extract more information such as the date in which the article was published, we believe that it would not be relevant for our system.

2.2 Text extraction and data storage

We used Python to extract the information from the XML documents once we have identified which documents are written in the language of our preference. For this purpose, we decided to use BeautifulSoup library in order to identify the tags and their respective content.

For the storage of this information we have employ Elasticsearch [1] which is a search engine based on Lucene. It will allow us to index our documents under the information retrieval ranking function Okapi BM25 [5] thus, reducing our development time.

3 Experiments

The organizers have been provided us with 30 topics following the structure:

- Keywords,
- conversational and,
- explanation

For every document indexed into our information retrieval system, we have provided: (1) id, (2) title, (3) relevant keywords and (4) document textual content. We have decided to structure our runs following the runs described in Table 1.

Table 1. Features associated with each run.

Run #	Run name	Components
1	sinai1	Search using keywords as query statement on fields: title, relevant keywords and document's textual content
2	sinai2	Search using conversational as query statement on fields: title, relevant keywords and document's textual content
3	sinai3	Search using explanation as query statement on fields: title, relevant keywords and document's textual content
4	sinai4	Search using keywords, conversational and explanation as query statement on fields: title, relevant keywords and document's textual content
5	sinai5	Search using keywords as query statement on fields: relevant keywords

4 Results

The organization provided us with a scoring for every each one of the run we submitted which features have been specified in Section 3. There has been another participant for this task and in order to compare scores the following statistics have been assessed:

1. **nDGC** or Normalized Discounted Cumulative Gain [2] is a non-binary relevant assessment of documents which are ranked in a retrieval result considering closeness to the ranking's top and bottom. The results obtained through the use of this measure can be analyzed graphically against the rest of the participants in Figure 1. We can observe that the marginal means are below 0.5 and the runs we have provided score below 0.3 which is significantly distant from rest of the participants.
2. **AP** [4] or Average Precision is a measure that combines recall and precision for ranked retrieval results for a given set of relevant documents against its expectation. As it happened with the normalized discounted cumulative gains, our runs do not score high for this measure indicating that we are not providing the right relevant documents.
3. **P@5** or Precision of the ranking's top 5 documents. With this measure have achieved better results at least for the very first of our runs which, as observed in Table 1, uses keywords as the query statement. Nonetheless, the rest of our entries seem to have been scored below 0.5 which again, indicates that not the appropriate relevant documents are being delivered.

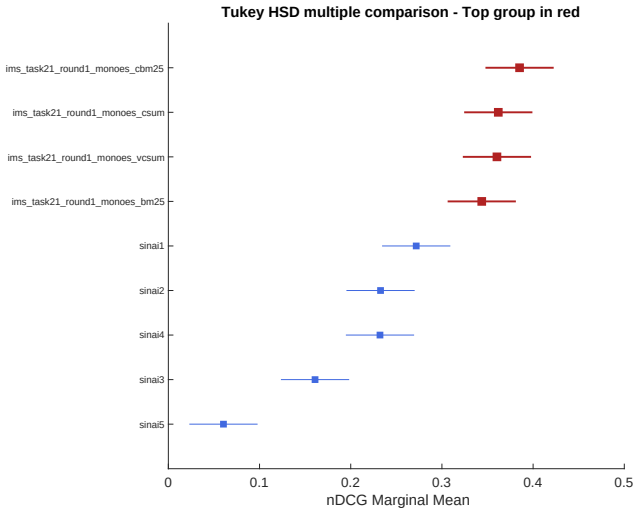

Fig. 1. Runs comparison using Normalized Discounted Cumulative Gain Marginal Mean.

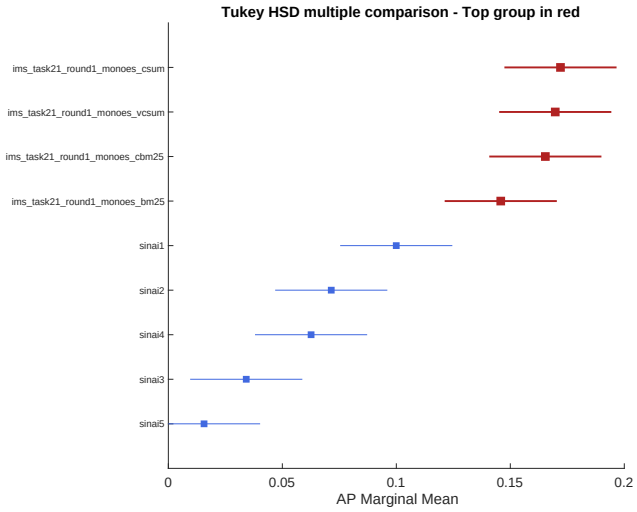

Fig. 2. Runs comparison using Average Precision Marginal Mean.

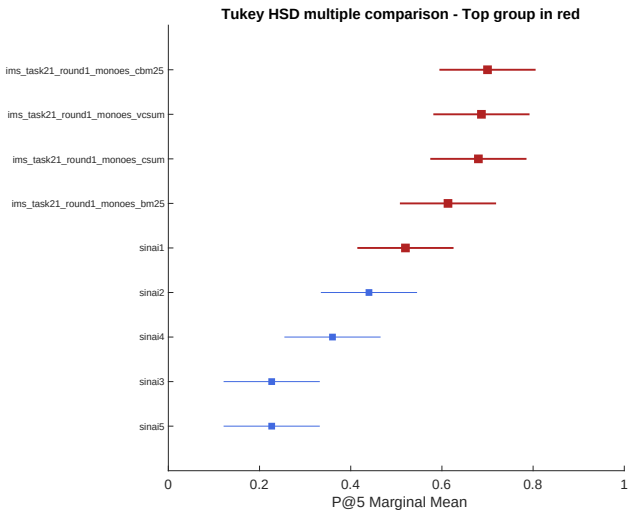

Fig. 3. Runs comparison using P@5 Marginal Mean.

4. For both **R and R-Prec**, measures for the proportion of the top-R retrieved relevant documents given a query, we find ourselves encountering the same situation proved within other measures.

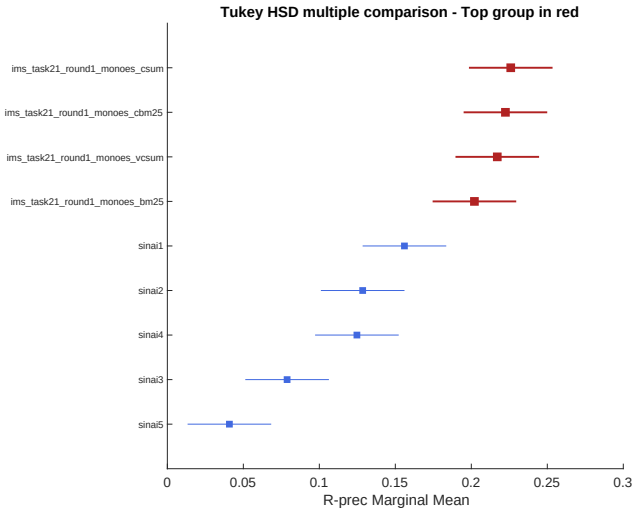

Fig. 4. Runs comparison using Marginal Mean and

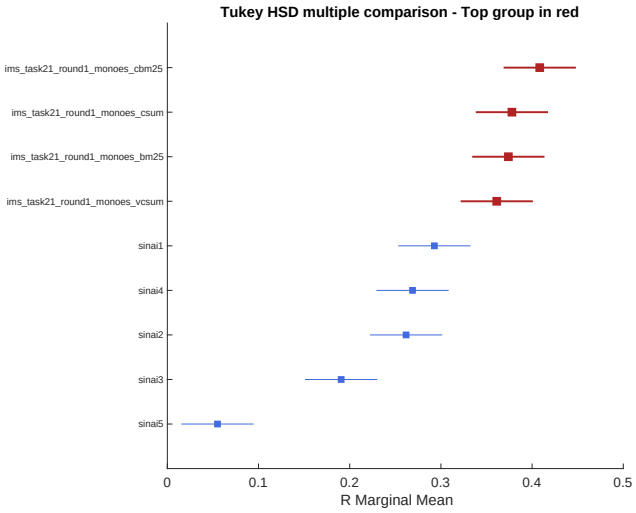

Fig. 5. Runs comparison using Marginal Mean and

5 Conclusion

In this paper we present our first participation at the MLIA-COVID-19 Workshop Task 2. We have taken a preliminary approach based on Lucene using the BM25 algorithm. In the future, we will aim to improve our system accuracy as well as to implement other information retrieval algorithms that may score higher.

References

[1] Gheorghe, R., Hinman, M.L., Russo, R.: Elasticsearch in action. Manning (2015)

[2] Järvelin, K., Price, S.L., Delcambre, L.M., Nielsen, M.L.: Discounted cumulated gain based evaluation of multiple-query ir sessions. In: European Conference on Information Retrieval, pp. 4–15, Springer (2008)

[3] López-Úbeda, P., Díaz-Galiano, M.C., Martín-Noguerol, T., Luna, A., Ureña-López, L.A., Martín-Valdivia, M.T.: Covid-19 detection in radiological text reports integrating entity recognition. Computers in Biology and Medicine **127**, 104066 (2020)

[4] Robertson, S.: A new interpretation of average precision. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 689–690 (2008)

[5] Robertson, S.E., Walker, S., Beaulieu, M.: Experimentation as a way of life: Okapi at trec. Information processing & management **36**(1), 95–108 (2000)

[6] Verspoor, K., Šuster, S., Otmakhova, Y., Mendis, S., Zhai, Z., Fang, B., Lau, J.H., Baldwin, T., Yepes, A.J., Martinez, D.: Covid-see: Scientific evidence explorer for covid-19 related research. arXiv preprint arXiv:2008.07880 (2020)

[7] Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W.R., Lo, K., Roberts, K., Soboroff, I., Wang, L.L.: Trec-covid: Constructing a pandemic information retrieval test collection. arXiv preprint arXiv:2005.04474 (2020)