

وطن-2004 و خليج-2004: مدونتان عربيتان للمعالجة الآلية للغات الطبيعية

Watan-2004 and Khaleej-2004: Two Arabic Corpora for Natural Language Processing tasks

مراد عباس

جامعة هواري بومدين للعلوم والتكنولوجيا، (الجزائر)، m.abbas@usthb.dz

تاريخ النشر: ديسمبر/2025

تاريخ القبول: 2025/12/01

تاريخ الإرسال: 2025/10/12

ملخص:

تعتبر البيانات نفط القرن الواحد والعشرين لما لها من بالغ الأهمية في تعلم الآلة. ونعرض في هذه الورقة وصفا لمجموعة بيانات، قمنا بإعدادها بشكل رئيسي لتصنيف النصوص العربية، ولأغراض أخرى تتعلق بمعالجة اللغات الطبيعية، مثل تلخيص النصوص، ونمذجة اللغة، والتعرف على الكيانات المسماة، بالإضافة إلى بناء موارد أخرى أو تطوير أدوات المعالجة الأولية المختلفة. تتكون مجموعة البيانات من مدونتين، هما وطن- 2004 وخليج - 2004، جُمعتا من صحف عربية إلكترونية. يبلغ حجمهما حوالي عشرة ملايين كلمة وثلاثة ملايين كلمة على التوالي.

الكلمات المفتاحية: المعالجة الآلية للغات الطبيعية، المدونات، وطن-2004، خليج-2004.

Abstract:

The dataset, we present in this paper, has been prepared mainly for Arabic text categorization. Nevertheless, it is useful for other purposes related to Natural Language Processing as text summarization, language modeling and Named entity recognition, as well as for building other resources or achieving different preprocessing tools. The dataset is composed of two Arabic corpora, namely Watan-2004 and Khaleej-2004, collected from online Arabic newspapers. Their sizes are around ten millions and three millions words respectively.

Keywords: Natural Language Processing, corpora, Watan-2004, Khaleej-2004.

1. Introduction

The Role of dataset in Natural Language Processing (NLP) is indispensable. They are not merely collections of text, they are the structured information, and labeled examples that allow building models capable of understanding, generating, and interacting with human language. The training data quality is the main key of the model strength. A number of points are considered in building such dataset. Size, diversity, representativeness of data are important parameters for the model to learn complex language nuances and represent the variations in language, style, and vocabulary.

It is clear that the famous powerful Large Language Models (LLMs) are data-intensive. However, their recent emergence has not made small-data machine learning irrelevant, because we often see them overkill or prohibitively expensive for practical, efficient, and specialized tasks, compared to small-data models. Therefore, developing relatively small data is still necessary for the reasons we mentioned above. The particular type of dataset that we are interested in is the one devoted to text categorization; which is basically a collection of texts/documents, each associated with a corresponding label.

In the following, we present in section 2 a brief description of the dataset, like the number of categories, and their size in terms of documents and words number. Then, we describe in section 3, the method of collecting the documents from the web and showing the different preprocessing steps and how representing the data. Finally, we focus on the value of both datasets by highlighting their impact on various research projects based on them.

2. Dataset

The source of our dataset is the articles downloaded in 2004 from two Arab online newspapers: AkhbarAlkhaleej and Alwatan¹. This was resulted in two corpora: Khaleej-2004 and Watan-2004. Watan-2004 is composed of more than 20.000 articles and covers six topics, namely: culture, religion, international news, local news, economy and sports (Table 1), while Khaleej-2004 is composed of 5690 articles and covers four topics, as indicated in Table 2.

¹<https://akhbar-alkhaleej.com/>
<https://alwatan.om/>

<i>Topic</i>	<i># documents / words</i>
Culture	2782 / 1.359.210
Religion	3860 / 3.122.565
Economy	3468 / 1.460.462
Local News	3596 / 1.555.635
International News	2035 / 855.945
Sports	4550 / 1.423.549
Total	20291 / 9.813.366

Table 1. Watan-2004 corpus statistics

<i>Topic</i>	<i>#documents / words</i>
International News	953 / 755.000
Local News	2398 / 893.000
Economy	909 / 578.000
Sports	1430 / 628.000
Total	5690 / 2.854.000

Table 2. Khaleej-2004 corpus statistics

3. Experimental Design, Materials and Methods

We used WinHTTrack to collect a large number of Web pages that contain newspapers articles in different topics. This Web crawler allows saving the html files in a well-organized way. We have developed a tool (using c++) to extract automatically the texts from the downloaded html files, and saved them in different categories (culture, religion, international news, local news, economy and sports).

Faced to thousands of resulted texts, we had to develop many tools related to text preprocessing required for text categorization as term frequencies computing (Table 3), light stemming, punctuation marks and stop words removal.

For Arabic, non-content words removal is not as simple as for English or French because of agglutination, which is one of the specificities of Arabic. For instance, the preposition “و” is not isolated from the following word. That is why we conducted light stemming to separate such non-content words and then to remove them.

Culture	Religion	Economy	Local	International	Sports
عربية 1867	الله 20233	في 3389	عام 2116	رئيس 2193	مباراة 2020
عام 1827	قال 5202	عام 2843	سلطنة 1875	في 2132	فريق 1795
كان 1773	صلى 4654	دول 2274	عمل 1755	عراق 2129	مركز 1792
عربي 1308	وسلم 4504	عمل 2117	وزارة 1524	وقال 2025	بطولة 1680
عالم 1192	كان 4427	سلطنة 1768	مهرجان 1438	عام 1817	منتخب 1676
عمل 1179	تعالى 4052	قطاع 1747	تم 1326	متحدة 1710	عام 1531
كانت 1122	رسول 3413	نقط 1646	فعاليات 1209	قوات 1494	اول 1499
فيلم 1103	ناس 3045	مليون 1555	مسقط 1189	كان 1314	اولمبية 1431
في 1078	قرآن 2600	سوق 1487	عامية 1105	حكومة 1277	ثاني 1197
كتاب 1063	إسلام 2211	شركة 1456	صحبة 1089	اميركية 1246	دورة 1195
فنية 1056	يكون 1884	ريال 1415	عمانية 1049	ماضي 1164	كرة 1148
فن 1053	نبي 1844	عماني 1317	تعليم 1043	ولايات 1107	اتحاد 1046
ثقافة 1042	دين 1675	شركات 1302	مركز 1036	مجلس 1084	ثانية 960

Table 3. Most frequent words in Watan-2004 corpus

4. Value of the Data

Khaleej-2004 and Watan-2004 are very useful for NLP research related to Arabic language: in topic detection [1-16], text categorization [27-33], text summarization [34-36], clustering [37, 38], document retrieval [39] and Named entity recognition [48-50]. This data has also been effectively utilized for building other resources [17-24], as well as for achieving quality assessment of Arabic corpora [25] and Wordnet Construction [26]. In addition, the dataset's utility extends across numerous advanced applications, including root extraction [40], exploring semantic relatedness [41], word sense disambiguation [42], sentiment analysis [43] and multi-word terms extraction [8, 44]. More fundamentally, it has also served as the cornerstone for building preprocessing tools [45,46] like stemming [47], tokenization and Part of speech tagging [48].

The dataset has proven critical for many other tasks as automatic vowelization of Arabic texts [52], embedded Arabic text detection and recognition in videos [53,54], identification of broadcast news story boundaries [55] and even Arabic dialects speech recognition systems [56].

Conclusion

In this paper, we presented in a very concise way, the dataset we built for NLP tasks, mainly for text categorization and topic identification. It is important to note that the Arabic language differs from other Latin languages in several aspects due to its features. Though advantageous in many situations, these features pose a hurdle that makes Arabic processing more complex.

Arabic's ability to group several semantic units into a single orthographic word results in a highly concise text representation, but also requires additional work in text processing field. Another characteristic of Arabic is the omission of short vowels (or diacritics), which is unusual for other languages. The consequence of the intentional omission of these vowels is that the word becomes more compressed and more ambiguous. All these challenges arose during the dataset processing, proving particularly difficult in text representation. The challenges we encountered involve preprocessing steps, such as calculating term frequency and removing stop words. The various works based on this dataset ² show its usefulness for the research community.

²A summary of dataset specifications is reported in Table 4.

References

- [1] Koulali, R., El-Haj, M., Meziane, A., Arabic Topic Detection using automatic text summarization, in *ACS International Conference on Computer Systems and Applications (AICCSA)*, Fes, 2013, pp. 1-4.
- [2] Naili, M., Chaibi, A., Ghézala, H., Arabic topic identification based on empirical studies of topic models, *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées*. 27 (2017) 45-59.
- [3] Alsaad, A., Abbod, M., Enhanced Topic Identification Algorithm for Arabic Corpora, in *17th UKSim-AMSS International Conference on Modelling and Simulation (UKSim)*, Cambridge, 2015, pp. 90-94.
- [4] Loukam, M., Hammouche, D., Belkredim, F.Z., Repérage de marqueurs pouvant conduire à l'extraction automatique de la thématique d'un texte en Arabe standard, in *Proceedings of the 8th National Conference on Computer Science, SNIB'2015*, Algeria, 2015, pp. 46-53.
- [5] Naili, M., Chaibi, A.H., Ghézala, H. B., Empirical study of LDA for Arabic topic identification, in *Proceedings of CARI, Tunisia*, 2016, pp. 138-145.
- [6] Abainia, K., Topic Identification of Noisy Texts: Statistical Approaches, *International Journal of Hidden Data Mining and Scientific Knowledge Discovery*, 1(1)(2015) 2-8.
- [7] Abainia, K., Ouamour, S., Sayoud, H., Neural Text Categorizer for topic identification of noisy Arabic Texts. In *IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, Marrakesh, 2015, pp. 1-8.
- [8] Koulali R., Meziane A., Topic Detection and Multi-word Terms Extraction for Arabic Unvowelized Documents, in: Salem M.V.M., Shaalan K., Oroumchian F., Shakery A., Khelalfa H. (eds) *Information Retrieval Technology. AIRS 2011. Lecture Notes in Computer Science*, vol 7097. Springer, Berlin, Heidelberg, 2011, pp. 614-623.
- [9] Koulali, R., Meziane, A., Experiments with Arabic topic detection, *Journal of Theoretical & Applied Information Technology*. 50(1) (2013) p28-32.
- [10] Koulali, R., Meziane, A., Feature Selection for Arabic Topic Detection using Named Entities, in *Proceedings of CITALA, Morocco*, 2014, pp. 243-246.
- [11] Abbas, M., Smaili, K., Berkani, D., Tr-classifier and knn evaluation for topic identification tasks. *International Journal on Information and Communication Technologies (IJICT)*, 3(3) (2010) 65-74.
- [12] Abbas, M., Smaili, K., Berkani, D., Multi-category support vector machines for identifying Arabic topics, *Advances in Computational Linguistics*, Special issue of *Journal of Research in computing Science*, 41(2009) 217-226.
- [13] Abbas, M., Smaili, K., Berkani, D., Evaluation of Topic Identification Methods on Arabic Corpora, *Journal of Digital Information Management*. 9(5) (2011) 185-192.
- [14] Abbas, M., Smaili, K., Berkani, D., Evaluation of Topic Identification Methods for Arabic Texts and their Combination by using a Corpus Extracted from the Omani Newspaper Alwatan, *Arab Gulf Journal of Scientific Research*. 29 (3/4) (2011) 183-191.
- [15] Abbas, M., Smaili, K., Berkani, D., Efficiency of TR-Classifer versus TFIDF, in *proceedings of International Conference on Integrated Intelligent Computing, ICIIC 2010*, India, 2010, pp. 233-237.
- [16] Abbas, M., Smaili, K., Comparison of Topic Identification Methods for Arabic Language, in *Recent Advances in Natural Language Processing (RANLP05)*, Borovets, 2005, pp. 14-17.
- [17] Al-Thubaity, A.O., A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3) (2015) 721-751.
- [18] El-Haj, M., Koulali, R., KALIMAT a multipurpose Arabic Corpus, in *Second Workshop on Arabic Corpus Linguistics (WACL-2)*, Lancaster, 2013, pp. 22-25.

- [19] Almeman, K., Lee, M., Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words, in proceedings of 1st international conference on Communications, signal processing, and their applications (ICSSPA), UAE, 2013, pp. 1-6.
- [20] Al-Thubaity, A., Khan, M., Al-Mazrui, M., Al-Mousa, M., New language resources for Arabic: corpus containing more than two million words and a corpus processing tool, in International Conference on Asian Language Processing (IALP), China, 2013, pp. 67-70.
- [21] El-khair, I.A., 1.5 billion words Arabic Corpus, 2016, arXiv preprint arXiv:1611.04033.
- [22] Ghoul, D., Ibrahim, A.H., Web Arabic corpus: Construction d'un large corpus arabe annoté morpho-syntaxiquement à partir du Web, in CEC-TAL'2015, Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, Tunisia, 2015, pp 12-16.
- [23] Sabbah, T., Selamat, A., A novel dataset for quranic words identification and authentication, in 2nd International Conference on Interactive Digital Media (ICIDM 2013), Malaysia, 2013, pp. 263-268.
- [24] Tarouti, A.L., ALI, F., LexBank: A Multilingual Lexical Resource for Low-Resource Languages (Doctoral dissertation, University of Colorado Colorado Springs. Kraemer Family Library), 2016.
- [25] Yahya, A., Salhi, A., Quality Assessment of General and Categorized Arabic Text Corpora, in 10th International Conference on Innovations in Information Technology, Tamil Nadu, 2014, pp. 149-158.
- [26] Al Tarouti, F., Kalita, J., Enhancing Automatic Wordnet Construction Using Word Embeddings, in Proceedings of the Workshop on Multilingual and Cross-lingual Methods in NLP, California, 2016, pp. 30-34.
- [27] Ghareb, A.S., Bakar, A.A., Hamdan, A.R., Hybrid feature selection based on enhanced genetic algorithm for text categorization, Expert Systems with Applications. 49 (2016) 31-47.
- [28] Mohammed I. Khaleel, Ismail I. Hmeidi, Hassan M. Najadat, An Automatic Text Classification System Based on Genetic Algorithm, in Proceedings of the 3rd Multidisciplinary International Social Networks Conference on Social Informatics 2016, Data Science 2016 (MISNC, SI, DS 2016). ACM, New York, New York, 2016, pp. 1-7.
- [29] Ghareb, A., Hamdan, A., Bakar, A., Integrating Noun-Based Feature Ranking and Selection Methods with Arabic Text Associative Classification Approach. Arabian Journal for Science & Engineering (Springer Science & Business Media BV), 39(11) (2014) 7807-7822.
- [30] Alahmadi, A., Automatic text classification using bag of words and bag of concepts based representations. Doctoral thesis, 2016.
- [31] Abbas, M., Identification de thèmes pour la reconnaissance automatique de la parole, Doctoral thesis, 2005.
- [32] Ayadi, R., Maraoui, M., Zrigui, M., A Survey of Arabic Text Representation and Classification Methods, Research in Computing Science (117) (2016) 51-62.
- [33] Dawoud, H. M. (2013). Combining different approaches to improve Arabic text documents classification (Doctoral dissertation, MSc Thesis, Islamic University).
- [34] Zidoum, H., Al-Maamari, A., Al-Amri, N., Al-Yahyai, A., Al-Ramadhani, S., Extracting Sentences Using Lexical Cohesion for Arabic Text Summarization. Int. J. Comput. Linguistics Appl. 6(1) (2015) 81-102.
- [35] Zidoum, H., Al-maamari, A., Al-Amri, N., Al-Yahyai, A., Al-Ramadhani, S., Computing Lexical Chains for Automatic Arabic Text Summarization. International Journal of Computing, Communication and Instrumentation Engineering (IJCCIE). 2(1) (2015) pp. 26-32.
- [36] Waheeb, S. A., Multi-document text summarization using text clustering for Arabic Language (Doctoral dissertation, Universiti Utara Malaysia), 2014.

- [37] Kelaiaia, A., Merouani, H.F., Clustering with probabilistic topic models on Arabic texts, in *Modeling Approaches and Algorithms for Advanced Computer Applications*, Springer International Publishing, 2013, pp. 65-74.
- [38] Kelaiaia, A., Merouani, H.F., Clustering with probabilistic topic models on Arabic texts: a comparative study of LDA and K-means. *Int. Arab J. Inf. Technol.*, 13(2) (2016) 332-338.
- [39] Al-Dayel, A., Ykhlef, M., Enhanced Arabic Document Retrieval Using Optimized Query Paraphrasing, *Arabian Journal for Science & Engineering* (Springer Science & Business Media BV), 40(11) (2015) pp. 3211–3232.
- [40] Alsaad, A., Enhanced root extraction and document classification algorithm for Arabic text (Doctoral dissertation, Brunel University London), 2016.
- [41] Toumouh, A., Widdows, D., Lehireche, A., Exploring Semantic Relatedness in Arabic Corpora using Paradigmatic and Syntagmatic Models. *International Journal of Information Engineering and Electronic Business*, 8(1) (2016) pp. 37-47.
- [42] Bakhouch, A., Yamina, T., Schwab, D., Tchekmedjiev, A., Ant colony algorithm for Arabic word sense disambiguation through English lexical information. *International Journal of Metadata, Semantics and Ontologies*, 10(3) (2015) 202-211.
- [43] Altowayan, A.A., Tao, L., Word embeddings for Arabic sentiment analysis, in *IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3820-3825.
- [44] Meghawry, S., Elkorany, A., Salah, A., Elghazaly, T., Semantic Extraction of Arabic Multiword Expressions. *Computer Science & Information Technology*. 5 (2015) 21-31.
- [45] Saad, M.K., The impact of text preprocessing and term weighting on Arabic text classification (Doctoral dissertation, The Islamic University-Gaza), (2010).
- [46] Sabri, T., Kissi, M., Bahassine, S., El Beggar, O., Analytics of ensemble learning-based methods for Arabic text classification, 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), Casablanca, Morocco, 2023, pp. 1-6.
- [47] El-Defrawy, M., El-Sonbaty, Y., Belal, N.A., Cbas: Context based Arabic stemmer. *arXiv preprint arXiv:1611.00027*, 2015.
- [48] Ahmed H. Aliwy, Tokenization as Preprocessing for Arabic Tagging System, *International Journal of Information and Education Technology*. 2 (4) (2012) 348-353.
- [49] Koulali, R., Meziane, A., A contribution to Arabic named entity recognition, in 10th International Conference on ICT and Knowledge Engineering, Bangkok, 2012, pp. 46-52.
- [50] Koulali, R., Meziane, A., A combined approach to Arabic named entity recognition using SVM and pattern extracted method applied to topic detection, in *proceedings of the International Conference on Arabic Language Processing*, Morocco, 2012, pp.131-137.
- [51] Mesmia, F.B., Friburger, N., Haddar, K., Maurel, D., Construction d'une cascade de transducteurs pour la reconnaissance des dates à partir d'un corpus Wikipédia, , in *CEC-TAL'2015, Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications*, Tunisia, 2015, pp 8-11.
- [52] Bebah, M., Amine, C., Azzeddine, M., Abdelhak, L., Hybrid approaches for automatic vowelization of Arabic texts, *arXiv preprint arXiv:1410.2646*, 2014.
- [53] Yousfi, S., Berrani, S. A., Garcia, C., Contribution of recurrent connectionist language models in improving LSTM-based Arabic text recognition in videos, *Pattern Recognition*. 64 (2017) 245-254.
- [54] Yousfi, S., Embedded Arabic text detection and recognition in videos (Doctoral dissertation, Université de Lyon), 2016.
- [55] Khalaf, Z.A., Ping, T.T., Automatic identification of broadcast news story boundaries using the unification method for popular nouns, in *Federated Conference on Computer Science and Information Systems (FedCSIS)*, Poland, 2013, pp. 577-584.
- [56] Almeman, K.A., Reducing out-of-vocabulary in morphology to improve the accuracy in Arabic dialects speech recognition (Doctoral dissertation, University of Birmingham), 2015.

Appendix

More specific subject area	<i>Computational Linguistics, Text Categorization, Topic Identification, Language Modeling, Modern Standard Arabic</i>
Type of data	<i>text file</i>
How data was acquired	<i>Html files were downloaded using HTTrack website copier, from two online Arabic newspapers Alwatan and Akhbar AlKhaleej.</i>
Data format	<i>Raw</i>
Experimental factors	<i>Texts were extracted from the downloaded html files, and then punctuation marks have been removed. Texts have been divided into a number of categories (topics).</i>
Experimental features	<i>Computing terms frequencies, extraction of non content words, Texts representation and automatic topic identification.</i>
Data accessibility	<i>The data is freely available at:</i> https://sourceforge.net/projects/arabiccorpus/files/

Table 4. Summary of dataset specifications.