ARC-WMI: Towards Building Arabic Readability Corpus for Written Medicine Information

Abeer Aldayel[1], Hend Al-Khalifa[2], Sinaa Alaqeel[3], Norah Abanmy[4], Maha Al-Yahya[5], Mona Diab[6]

[1,2,5]College of Computer and Information Science and [3,4]College of Pharmacy
King Saud University Riyadh, Saudi Arabia
{ [1]aabeer|[2]hendk|[3]salageel|[4]nabanmy|[5]malyahya }@ksu.edu.sa

[6]Department of Computer Science, the George Washington University
[6]mtdiab@gwu.edu

Abstract

Developing easy-to-read written medicine information continues to be a challenge in health communication. Readability aims to gauge the difficulty level of a text. Various formulas and machine learning algorithms have proposed to judge the readability of health materials and assist writers in identifying possible problems related to text difficulty. For this reason, having corpus annotated with readability levels is fundamental to evaluating the readability formulas and training machine learning algorithms. Arabic suffers from a lack of annotated corpora to evaluate text readability, especially for health materials. To address this shortage, we describe a baseline results towards constructing readability corpus ARC-WMI, a new Arabic collection of written medicine information annotated with readability levels. We compiled a corpus of 4476 sentences with over 61k words, extracted from 94 sources of Arabic written medicine information. These sentences were manually annotated and assigned a readability level ("Easy," "Intermediate," or "Difficult") by a panel of five health-care professionals.

Keywords: Corpus annotation, Readability corpus, Written medicine information

1. Introduction

Corpus annotation is the practice of adding metadata to a collection of text (Baker, 1997). These metadata relate to specific parts of the text (i.e., a word or a sentence) and are used to add both linguistic and descriptive information to it. Annotated corpora emerged to model various language phenomena and to train algorithms (Pustejovsky and Stubbs, 2012). There are different types of annotation tasks: one relates to linguistic models, such as a semantic annotated corpus (Basile et al., 2012) and a syntax annotated corpus (Brants et al., 2002), and the other relates to natural language processing (NLP) tasks, such as an inference corpus (Bowman et al., 2015).

Many recent studies have emerged to address the need for large health materials corpora with linguistic or NLP related metadata added to the text. There are health materials annotated corpora with sentiment-related information, such as clinical sentiment corpus (Deng et al., 2016), and other health related materials annotated with linguistics metadata, such as clinical part-of-speech tagging corpus (Pakhomov et al., 2006). Adding these metadata to health materials provides better insight into the data and facilitates obtaining robust results from the analyses.

Written medicine information (WMI) refers to the written information leaflet that accompanies medications (Koo et al., 2006). WMIs play an important role in educating consumers about their medicines. To contribute effectively to healthcare decision-making, these resources should be written at a level readable by any patient. Known as health literacy, measuring the readability for health related text is a long-established problem. Health literacy is defined as the degree to which individuals have the ability to understand basic health to make appropriate health decisions (Hewitt, 2012). Different methods have been used, such as traditional formulas and machine learning algorithms, to predict the text difficulty level and automatically predict the level of text readability. These methods need to be evaluated using a corpus annotated with readability levels (Koo et al., 2006).

In this paper, we introduce the ARC-WMI Arabic Readability Corpus. Comprised of more than 4000 sentences, it contains WMIs annotated with readability levels and collected from two sources: the Saudi Food and Drug Authority (SFDA)[1] and the King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE).[2] The ARC-WMI will address the need for a readability corpus to evaluate the readability algorithm and the formulas in the Arabic health domain.
This paper is organized as follows: Section 2 reviews the related work on the readability corpus field. Section 3 outlines the constructed corpus. Section 4 presents the methodology (in detail) that we used for the annotation process. The conclusion and future directions follow in Section 5.

[1] http://www.sfda.gov.sa/En/Pages/default.aspx

[2] https://www.kaahe.org/en/

2. Related works

Numerous studies have emerged to address the need for a gold standard corpus for readability assessments in the health field. There are two common methods for evaluating text readability: (1) direct evaluation and (2) pair-wise comparison. In the direct evaluation method, the annotator assigns absolute scores or labels that reflect the text difficulty and uses the resulting mean readability score as the overall text difficulty score. Many studies, such as the one by (Kandula and Zeng-Treitler, 2008), where they annotated 324 health documents with the readability level based on a 1–7 Likert scale, follow this method to annotate text for readability. Another study (Rosemblat et al., 2006), used the same method to evaluate the readability of 22 consumer health texts based on linguistic and stylistic features. In the pairwise comparison method, the annotator will compare between two texts and judge the relative readability score between them. However, in this study (Van Oosten and Hoste, 2011), they used a pair-wise comparison to evaluate the readability of a large corpus that contained domain-specific documents, manuals, and patient information leaflets.

There has been a significant amount of work on linguistic related corpora for Arabic text, including morphological segmentation (Dukes and Habash, 2010), punctuated corpora (Zaghouani and Awad, 2016), and in-depth work on sentiment corpora (Abdul-Mageed and Diab, 2012). In contrast, Arabic suffers from a shortage of well-formed readability corpus, especially for health related materials. In this paper, we construct a collection of readability annotated WMIs texts to describe an ongoing effort to fill this gap.

3. Corpus description

In the Arabic Readability Corpus for Written Medicine Information (ARC-WMI), the readability annotation was conducted at the sentence level in which selected sentences from each piece of WMI was evaluated based on three readability levels (Easy, Intermediate, and Difficult). A total of 4476 sentences and approximately 61k words were collected from 94 WMIs. The WMIs were collected from two sources: 47 WMI from the Saudi Food and Drug Authority (SFDA) and 47 WMI from the King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE) . Table 2 illustrates the word and sentence distributions in each source. These two sources have different text structures and use different subheadings and sections, as shown in Figure 2, which forced us to define the distribution of the sentences for each WMI. In our corpus, we designed a coding scheme to enable the unique and descriptive tag identification for the sentences of any of the WMIs**.**

The ID tag naming follows this pattern: “$Source(file_n)\_S_z\_(Sentence_t)$”, where S_z indicates the section number for each destination source (KAAHE or SFDA), as shown in Table 1. We defined the annotation values for the tags as integer numbers related to the text difficulty level (1. Easy; 2. Intermediate; 3. Difficult).

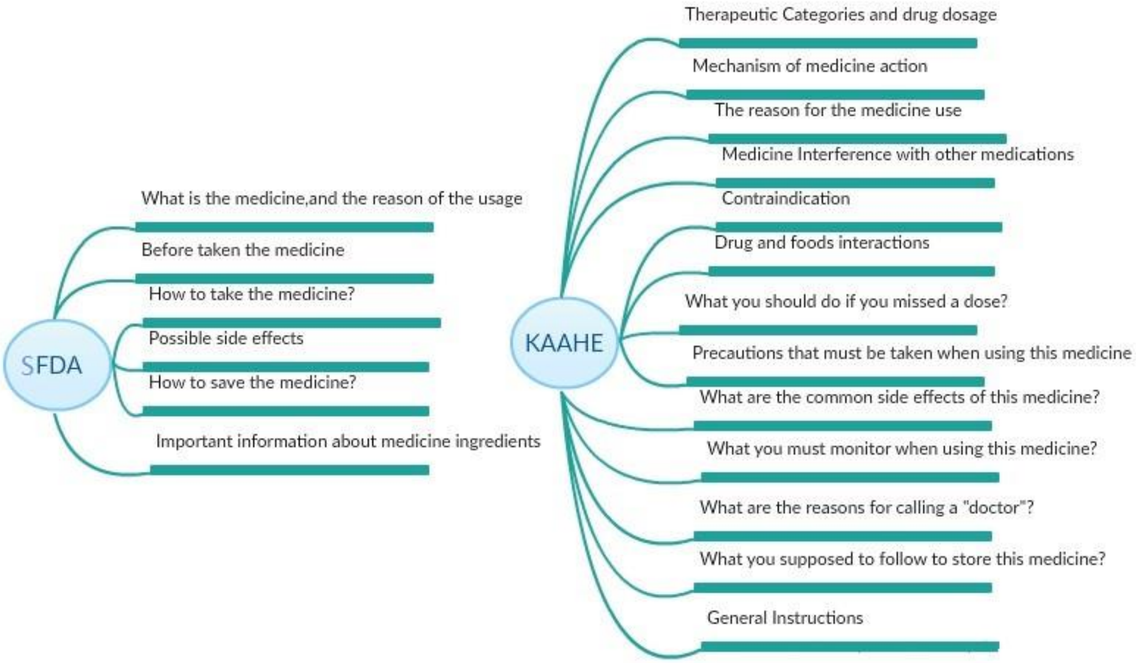
SFDA
What is the medicine,and the reason of the usage
Before taken the medicine
How to take the medicine?
Possible side effects
How to save the medicine?
Important information about medicine ingredients
KAAHE
Therapeutic Categories and drug dosage
Mechanism of medicine action
The reason for the medicine use
Medicine Interference with other medications
Contraindication
Drug and foods interactions
What you should do if you missed a dose?
Precautions that must be taken when using this medicine
What are the common side effects of this medicine?
What you must monitor when using this medicine?
What are the reasons for calling a "doctor"?
What you supposed to follow to store this medicine?
General Instructions

Figure 1 SFDA and KAAHE Structure

Table 1 Sentence coding

SFDA section (English)	**SFDA section (Arabic)**	**Sentence coding**
What is the medicine, and the reason of the usage?	ما هو الدواء و ما هي دواعي استعماله	SFDA(File#)_S0_(Sentence#)
Before taken the medicine	قبل القيام بتناول الدواء	SFDA(File#)_S1_(Sentence#)
How to take the medicine?	كيف تتناول الدواء	SFDA(File#)_S2_(Sentence#)
Possible side effects	الآثار الجانبية المحتملة	SFDA(File#)_S3_(Sentence#)
How to save the medicine?	كيف تقوم بحفظ الدواء	SFDA(File#)_S4_(Sentence#)
Important information about medicine ingredients.	معلومات مهمة حول بعض مكونات الدواء	SFDA(File#)_S5_(Sentence#)
KAAHE section (English)	**KAAHE section (Arabic)**	**Sentence coding**
Therapeutic Categories and drug dosage and	التَّصنيفُ العِلاجيُّ للدَّواء والجرعةُ الدَّوائيَّة	KAAHE(File#)_S0_(Sentence#)
mechanism of action	آليَّةُ عَمَل الدَّواء	KAAHE(File#)_S1_(Sentence#)
The reason for the medicine use	دَواعي استِعمال الدَّواء	KAAHE(File#)_S2_(Sentence#)
contraindication	مَوانِعُ استِعمال الدَّواء	KAAHE(File#)_S3_(Sentence#)
Best way of taking medicine	ما هي الطَّريقةُ المُثلى لاستِعمال هذا الدَّواء؟	KAAHE(File#)_S4_(Sentence#)
Drug and foods interactions	تَداخلُ الدَّواء مع الطَّعام	KAAHE(File#)_S5_(Sentence#)
Medicine Interference with other medications	تَداخلُ الدَّواء مع الأَدوية الأخرى	KAAHE(File#)_S6_(Sentence#)
What to do if you missed a dose?	ماذا أفعلُ إذا تأخَّرتُ عن موعد إحدى الجرعات؟	KAAHE(File#)_S7_(Sentence#)
Precautions that must be taken when using this medicine	ما هي الاحتياطاتُ التي يجب مُراعاتُها لدى استِعمال هذا الدَّواء؟	KAAHE(File#)_S8_(Sentence#)
What are the common side effects of this medicine?	ما هي التأثيراتُ الجانبيَّة الشَّائعة لهذا الدَّواء؟	KAAHE(File#)_S9_(Sentence#)
What you must monitor when using this medicine?	ماذا يجب على المَرء مراقبتُه عندَ استِعمال هذا الدَّواء؟	KAAHE(File#)_S10_(Sentence#)
What are the reasons for calling the health care resource "doctor"?	ما هي الأسبابُ التي تدعو لاستدعاء مورد الرعاية الصحِّية " الطَّبيب " على الفور؟	KAAHE(File#)_S11_(Sentence#)
What you supposed to follow when store this medicine?	ما المَفروضُ اتِّباعُه لدى تَخزين هذا الدَّواء؟	KAAHE(File#)_S12_(Sentence#)
General Instructions	إرشاداتٌ عامَّة	KAAHE(File#)_S13_(Sentence#)

Table 2 Words and sentences distribution

Source	**SFDA**	**KAAHE**	**Total**
Word count	31995	29400	61395
Sentences	2231	2245	4476

4. Annotation Methodology

We followed the annotation process pipeline defined by (Pustejovsky and Stubbs, 2012) to create an ARC-WMI with a readability annotations. Figure 2 shows the workflow and the main phases for the readability annotation process. The annotation modelling and guidelines define the annotation policy for the annotators and they identify the annotation values to be assigned for each sentence. Our readability guidelines focused on identifying the readability levels and they described how the difficulty level should be assigned for a given sentence. We defined three levels of text readability (Easy, Intermediate, and Difficult), as shown in Table 4.

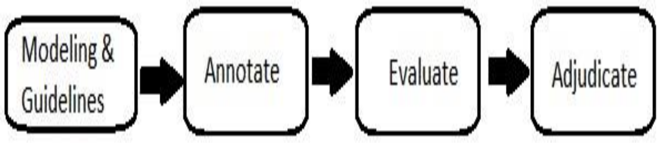

Figure 2 Annotation process pipeline

These levels were derived from a study by (Leroy et al., 2008), which evaluates the sentence based on the vocabulary used, the syntax structure, and the overall understanding. In the annotation phase, the sentence was judged for its readability by five health-care professionals' annotators. Each sentence was evaluated by two annotators to ensure the overlap between the annotation values.

In addition, these expert annotators have a pharmacy education background to ensure they can easily interpret and follow (Leory et al. ,2008) health information evaluation criteria. Each sentence was evaluated as "Easy" "Intermediate," or "Difficult," where readability was defined as a subjective judgment of how easily a reader could extract the information from the WMI.

In the evaluation phase, we measured the efficiency of the annotation's results using the Inter-Annotator Agreement (IAA) score. We calculated the IAA using kappa statistics for comparing two annotations against each other, based on Landis and Koch guidelines (Landis and Koch, 1977), to interpret the kappa value and define the agreement level.

Table 3 Example annotations from the corpus

Sentence ID	Sentence	English translation	annotator 1	annotator 2
KAAHE37_S0_3	أمَّا جرعةُ الصِّيانَة فهي 25-100 ملغ/اليوم على دفعة أو دفعتين بعدَ أسبوعين من بَدء العِلاج عندَ الضَّرورة؛ ويمكن إضافةُ المدرَّات حسب الحاجة.	The maintenance dose is 25-100 mg/ day on a batch or two batches after two weeks of starting treatment, it is possible to add diuretics when necessary	3	2
KAAHE36_S12_45	يُحفَظ الدَّواءُ في درجة حرارة الغرفة.	Keep the medicine at room temperature	1	1
KAAHE11_S3_11	إذا كان المَريضُ يعانِي من أمراض الكبد أو من إسهال شَديد يُسمَّى التهاب القولون الغِشائي الكاذب.	If the patient is suffering from liver disease or from severe diarrhea called pseudo-colitis.	2	1

Table 4 Annotation guideline

Readability level	Definition
Easy	Contains small number of medical vocabulary and syntax structure used by the average consumer and he/she can understand the sentence without any help.
Intermediate	Contains medical vocabulary and syntax structure used in consumer health education and he/she can understand the sentence as consumer health education.
Difficult	Contains many medical vocabulary and syntax structure used by health professionals. Only health professionals can understand the sentence.

Table 5 shows the resulting IAA with average Inter-Annotator agreement 22% for the complete annotated dataset. This result indicates a fair agreement level with noticeable fluctuation in the agreement levels between the annotators. To resolve the conflict we used a third party judge to settle the differences in the annotation set. Table 3 presents sample of annotations values from dataset. Considering that the guideline definitions were derived from (Leory et al. 2008), still the annotators find it difficult to distinguish between Intermediate and Easy sentences. In addition the annotators tend to choose in case of uncertainty the intermediate level.

Adjudicating was conducted to resolve the conflicts between the annotators' results. In the annotation set, the differences between the annotations occurred because the text readability is based on the annotator's intuition to evaluate the difficulty level of the text. The differences between the annotations are expected and they are legitimate, based on the nature of the readability process (Finlayson, 2011). To finalize the results of the annotation, an adjudicator was employed to compare between the annotation values and to resolve any conflicts between the annotator's assigned values, to produce the final version of the annotated corpus. Table 6 shows the distribution of the sentences for each category, with an average sentence length of 15 for Easy, 21 for Intermediate, and 25 for Difficult.

Table 5 The resulted Inter-Annotator Agreement (IAA) score

Annotator Pair\agreement	IAA (Kappa)
Annotator (1 & 2)	0.48
Annotator (2 & 3)	0.11
Annotator (3 & 4)	0.028
Annotator (4 & 5)	0.28

Table 6 Distribution of sentences

	Easy	Intermediate	Difficult
Sentences count	3224	918	334
Words count (per sentence)	38501	15815	7079
AVG sentence length	15	21	25

5. Conclusion and Future Directions

In this paper, we presented the ARC-WMI Readability Corpus for WMIs, which is the first computationally analyzed Arabic corpus for readability assessment for the health domain. This corpus contains over 61k words and 4476 sentences annotated with three readability levels (Easy, Intermediate, and Difficult). We believe that a readability annotated corpus would be extremely valuable for future developments in computational readability research, especially for health literacy studies. Future work includes a further extension of the corpus along with guideline enhancement to improve the overall IAA results. The IAA values, can be improved by the experience gained over time by the annotators during the annotation process and by updating the annotation guidelines to simplify the readability assessment criteria and include a clear criteria for uncertainty cases as well. Finally, we will soon release the preliminary version of ARC-WMI corpus [3] under a Creative Commons License, so the research community can benefit from it.

Acknowledgment

This Project was funded by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (INF 2822).

References

Abdul-Mageed, M., Diab, M.T., 2012. AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis. Presented at the LREC, Citeseer, pp. 3907–3914.

Baker, J.P., 1997. Consistency and accuracy in correcting automatically tagged data. Garside et al.(1997) 243–250.

Basile, V., Bos, J., Evang, K., Venhuizen, N., 2012. Developing a large semantically annotated corpus. Presented at the LREC, pp. 3196–3200.

Bowman, S.R., Angeli, G., Potts, C., Manning, C.D., 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.

Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G., 2002. The TIGER treebank. Presented at the Proceedings of the workshop on treebanks and linguistic theories.

Deng, Y., Declerck, T., Lendvai, P., Denecke, K., 2016. The Generation of a Corpus for Clinical Sentiment Analysis. Presented at the International Semantic Web Conference, Springer, pp. 311–324.

Dukes, K., Habash, N., 2010. Morphological Annotation of Quranic Arabic. Presented at the LREC.

Finlayson, M.A., 2011. The Story Workbench: An Extensible Semi-Automatic Text Annotation Tool. Presented at the Intelligent Narrative Technologies.

Hewitt, M., 2012. Facilitating State Health Exchange Communication Through the Use of Health Literate Practices: Workshop Summary. National Academies Press.

Kandula, S., Zeng-Treitler, Q., 2008. Creating a gold standard for the readability measurement of health texts. Presented at the AMIA.

Koo, M., Krass, I., Aslani, P., 2006. Enhancing patient education about medicines: factors influencing reading and seeking of written medicine information. Health Expectations 9, 174–187.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics 159–174.

Leroy, G., Miller, T., Rosemblat, G., Browne, A., 2008. A balanced approach to health information evaluation: A vocabulary-based naïve Bayes classifier and readability formulas. Journal of the American Society for Information Science and Technology 59, 1409–1419.

Pakhomov, S.V., Coden, A., Chute, C.G., 2006. Developing a corpus of clinical notes manually annotated for

[3] https://github.com/iwan-rg/ARC-WMI

part-of-speech. International journal of medical informatics 75, 418–429.

Pustejovsky, J., Stubbs, A., 2012. Natural language annotation for machine learning. O'Reilly Media, Inc.

Rosemblat, G., Logan, R., Tse, T., Graham, L., 2006. Text features and readability: expert evaluation of consumer health text. Presented at the Mednet 2006: 11th World Congress on Internet in Medicine the Society for Internet in Medicine, Citeseer.

Van Oosten, P., Hoste, V., 2011. Readability annotation: Replacing the expert by the crowd. Presented at the Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, pp. 120–129.

Zaghouani, W., Awad, D., 2016. Building an Arabic Punctuated Corpus. Presented at the Qatar Foundation Annual Research Conference Proceedings, HBKU Press Qatar, p. SSHAPP3148.