8-1-2024

A comprehensive dataset for Arabic word sense disambiguation

Sanaa Kaddoura
Zayed University, sanaa.kaddoura@zu.ac.ae

Reem Nassar
Zayed University

Data in Brief 55 (2024) 110591

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A comprehensive dataset for Arabic word sense disambiguation

Sanaa Kaddoura*, Reem Nassar

Computing and Applied Technology, College of Technological Innovation, Zayed University, UAE

ARTICLE INFO

Article history:
Received 22 January 2024
Revised 3 May 2024
Accepted 30 May 2024
Available online 4 June 2024

Dataset link: Dataset for Arabic Word Sense Disambiguation (Original data)

Keywords:
Labelled data
Word sense disambiguation
Machine learning
Deep learning
GPT3.5
Natural language processing
Arabic language

ABSTRACT

This data paper introduces a comprehensive dataset tailored for word sense disambiguation tasks, explicitly focusing on a hundred polysemous words frequently employed in Modern Standard Arabic. The dataset encompasses a diverse set of senses for each word, ranging from 3 to 8, resulting in 367 unique senses. Each word sense is accompanied by contextual sentences comprising ten sentence examples that feature the polysemous word in various contexts. The data collection resulted in a dataset of 3670 samples. Significantly, the dataset is in Arabic, which is known for its rich morphology, complex syntax, and extensive polysemy. The data was meticulously collected from various web sources, spanning news, medicine, finance, and more domains. This inclusivity ensures the dataset's applicability across diverse fields, positioning it as a pivotal resource for Arabic Natural Language Processing (NLP) applications. The data collection timeframe spans from the first of April 2023 to the first of May 2023. The dataset provides comprehensive model learning by including all senses for a frequently used Arabic polysemous term, even rare senses that are infrequently used in real-world contexts, thereby mitigating biases. The dataset comprises synthetic sentences generated by GPT3.5-turbo, addressing instances where rare senses lack sufficient real-world data. The dataset collection process involved initial web scraping, followed by manual sorting to distinguish word senses, supplemented by thorough searches by a human expert to fill in missing contextual sentences. Finally, in

* Corresponding author.
E-mail address: sanaa.kaddoura@zu.ac.ae (S. Kaddoura).
Social media: @sanaa_kaddoura (S. Kaddoura), @Reem62434634 (R. Nassar)

https://doi.org/10.1016/j.dib.2024.110591

instances where online data for rare word senses was lacking or insufficient, synthetic samples were generated. Beyond its primary utility in word sense disambiguation, this dataset holds considerable value for scientists and researchers across various domains, extending its relevance to sentiment analysis applications.

Specifications Table

Subject	Data Science
Specific subject area	The dataset comprises Arabic sentences designed for natural language processing classification tasks, with labeled senses for polysemous words, making it particularly valuable for Word Sense Disambiguation models. It is well-suited for the application of machine learning techniques.
Type of data	Raw Data, Table
Data collection	The collection process began with selecting polysemous terms and extracting their various senses from the Arabic dictionary, resulting in a dataset including 367 senses. Then, ten example sentences were collected for each sense, illustrating the polysemous word within its corresponding context. The sentences were curated from diverse sources, including arts, political, medical, meteorological, economic, etc. The dataset also incorporates thoroughly generated sentences by GPT3.5-turbo. Finally, the dataset was reviewed by a native Arabic speaker to ensure accuracy and linguistic fidelity in representing the various senses of the polysemous terms.
Data source location	The dataset includes sentences from articles published from different locations and different domains such as political and medical. The main website used for polysemous words and content collection is Wikipedia.
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/pmdbs9tby8.1 Direct URL to data: https://data.mendeley.com/datasets/pmdbs9tby8/1 Instructions for accessing these data: There are no specific instructions.
Related research article	Kaddoura, S., & Nassar, R. (2024). EnhancedBERT: A Feature-rich Ensemble Model for Arabic Word Sense Disambiguation with Statistical Analysis and Optimized Data Collection. Journal of King Saud University-Computer and Information Sciences, 101,911. https://doi.org/10.1016/j.jksuci.2023.101911

1. Value of the Data

- This dataset was created due to the absence of publicly available resources for Arabic Word Sense Disambiguation (WSD), where research lags behind that in other languages [1]. While descriptions of datasets may exist in the literature, the inaccessibility of downloads restricts their usefulness. This dataset fills a gap in the availability of linguistic resources for Arabic Natural Language Processing (NLP) tasks. It uniquely contributes to WSD by focusing on the Arabic language, which is known for its rich morphology, complex syntax, and extensive polysemy, which represents challenges for NLP research. Therefore, this dataset will advance the research in the Arabic Language.
- This dataset offers researchers and practitioners many applications, serving as a standardized benchmark for evaluating and advancing WSD techniques in the Arabic language domain. Additionally, it is a valuable resource for WSD knowledge-based algorithms, functioning as an enriched dictionary for enhanced algorithmic performance. It fills a critical gap in the availability of resources tailored to the Arabic language, enabling researchers to develop and evaluate NLP algorithms that are optimized for Arabic text.
- This dataset can improve Arabic NLP applications, including sentiment analysis and machine learning-based text classification.

- This dataset is not limited to a single domain, so it can be applied to discern specific senses across diverse fields. Its generality is crucial as a single Arabic word may assume varied senses in distinct domains, such as a proper noun, denoting a disease, referencing a religious concept, or holding financial significance. Neglecting any domain risks ignoring or missing specific meanings, leading to the construction of biased models.
- This dataset includes examples for a single sense with varying sentence lengths, domains, and word positions, which is crucial for building a robust NLP algorithm.
- Researchers can also use this dataset to test their models' generalization capabilities with diverse datasets.

2. Background

The dataset has proved its effectiveness in model learning and thus effectively disambiguated Arabic words in the original article [2], representing a significant advancement over prior Arabic WSD research. It has also proved its effectiveness in improving the performance of downstream tasks like sentiment analysis and can be used for other tasks like spam detection [3] and machine translation [4]. Previously, authors kept their WSD datasets private, revealing only sense statistics in their papers. The dataset by El-Razzaz et al. [5] offered initial public data into the Arabic WSD domain, but it exhibited inconsistencies, with some sense examples missing or incomplete. The article [2] recognizes the critical role of comprehensive data for a robust WSD system. Arabic NLP encounters challenges due to linguistic complexities. Modern Standard Arabic (MSA) tokens exhibit an average degree of ambiguity at 19.2 %, significantly higher than the 2.3 % in most languages [6]. The absence of diacritics in most Arabic texts increases polysemy. This intricacy adversely impacts NLP model performance, emphasizing the need for a proficient WSD model to enhance the accuracy of NLP algorithms. Therefore, creating a comprehensive and diverse Arabic WSD dataset is fundamental to advancing Arabic language research.

3. Data Description

The dataset [7] is structured as a table stored in an Excel file, encompassing seven columns outlined in Table 1. The columns for the training of machine learning models include "Word," "Meaning," "POS," and "Sentence." These attributes are crucial in imparting the necessary information for the model to learn and make accurate predictions. Additionally, the columns "ID" and "Target ID" offer the opportunity for strategic data partitioning through stratified splitting. Leveraging these columns facilitates the division of the dataset based on specific criteria. Specifically, utilizing "ID" and "Target ID" allows for a refined approach, where the data can be split by the word itself and the sense associated with each instance. For instance, in the context of disambiguating a single word, the data can be extracted, and subsequently, each sense can be systematically divided into training and testing subsets.

Table 1
Description of each column in the dataset.

Column Name	Description
ID	A numerical value that represents the ID of a given polysemous word.
Word	The polysemous target word.
Meaning	The sense of the given polysemous word.
POS	The part of speech tag of the polysemous word at the given sense.
Sentence	An example sentence that includes the polysemous word within its sense context.
Target ID	Indicates the label corresponding to a Sentence with a target Sense.
Source	The source where the sentence is collected.

Table 2
Samples from the dataset.

Word	Sense	Sense translation	Example sentence	Sentence translation
حب	ميل النفس مع العقل، فإذا تجاوز العقل فهو العشق.	Love	عندما نكتب قصة **حب** جميلة لا يسأل النص الروائي عن صدقه وعن إمكاناته الإيصالية.	When we write a beautiful **love** story, the narrative does not inquire about its truthfulness or its communicative capabilities.
حب	جمع حَبّة.	Seeds or grains	وفي الماضي كان **حب** القمح يدق باليد ويستخدم في دقه ما يسمى بالمنحاز أو الياون أو الهاون.	In the past, wheat **grains** were threshed manually, using a using tool such as sickles, scythes, and flails.
حب	بثور.	Acne	البثور والرؤوس السوداء هي أنواع من **حب** الشباب التي تصيب الكثير من الناس.	Pimples and blackheads are types of **acne** that affect many people.

Table 2 shows three examples from the dataset for the polysemous word "حب." The table comprises the Arabic word, its sense, its translated sense, the Arabic sentence, and its translation. The polysemous word in the sentence is shown in bold. The table shows that "حب" has three different senses, each related to a specific category. The sense of love is related to emotions and relationships, thus forming the psychology category. The sense seeds or grains refer to agriculture and thus to the food category. The sense of Acne is related to dermatology and thus related to the health category.

Using a single category in Arabic will transform a polysemous word into one with a single sense. For example, the word "حب" presented in Table 2 carries only a single sense if one has limited the WSD study to medical purposes. However, this will introduce biases, and if the term "حب" comes in a different context, the model will fail to disambiguate it.

Table 3 illustrates the data distribution across various domains, presenting the number of examples extracted per domain. The table distinctly highlights the diverse data distribution across different categories, determined by the presence of data and examples within the specified sense context.

The dataset comprises 3670 samples, as evidenced by the statistics in Table 3, which are examples of sense examples. These samples are grouped into 367 senses for 100 polysemous terms, with ten samples per sense. The selection of these terms followed specific criteria to enhance the dataset's relevance and utility for Arabic WSD research. The primary criterion for word selection was the presence of multiple senses for each word and its frequent use in

Table 3
Number of samples taken from each domain.

Category	Number of sense examples
General	1311
Psychology	1046
Medical and Health	291
Religion	249
Political	138
Economics	111
Astronomy	97
Science and Education	69
Meteorology	64
Food and Culture	63
Sports	55
Arts and Fashion	54
People and Self	32
Geography	25
Agriculture	25
Mathematics and Abstractions	20
Labor	20

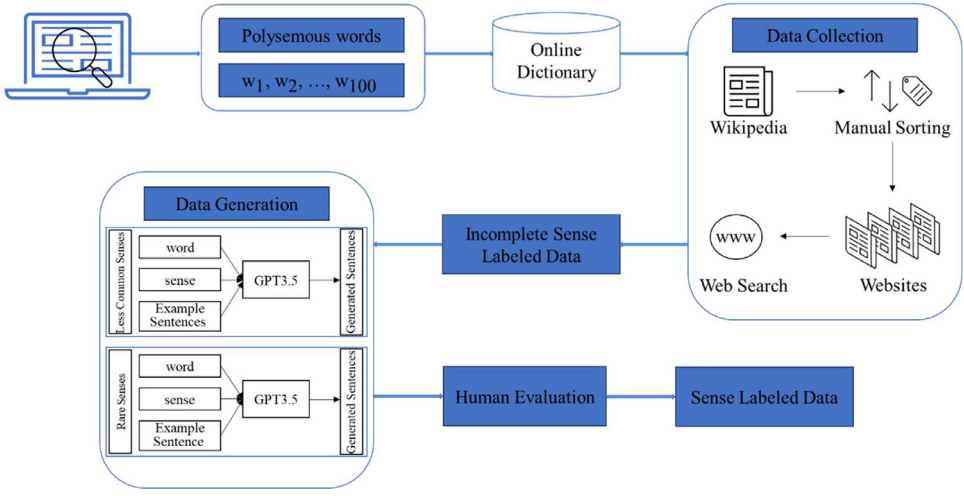

Fig. 1. Data collection procedure.

Arabic articles. This ensured that the dataset would provide broad opportunities for evaluating the effectiveness of WSD algorithms in disambiguating between different meanings within the context of Arabic text. While selecting the words, there was a focus on words that have diverse and representative contextual examples in Arabic sources available online particularly Wikipedia. The dataset's aim was to ensure its comprehensive representativeness by including words from different semantic domains and linguistic complexities. The goal is to provide a comprehensive resource that reflects the complexities of Arabic language usage by ensuring that the chosen words have practical significance for NLP tasks.

4. Experimental Design, Materials and Methods

This section describes the data collection process, which is described in Fig. 1. The data were collected from various domains like Wikipedia Foundation [8], Al-Jazeera [9], Al-Jazeera Documentary [10], Altibbi [11], Arabia Weather [12], Shifaa [13], Arabia CNN [14], Arabia BBC [15], Arabic Post [16], and Argaam [17]. The initial step entailed identifying a hundred polysemous words in Arabic, achieved through web surfing by a native Arabic speaker, focusing on frequently occurring terms across different articles. After identifying the polysemous terms, their different senses were assigned from online dictionaries like Almaany [18] and the Riyadh Dictionary of Contemporary Arabic Language [19]. The selection process for the hundred polysemous words in the dataset involved multiple stages, given the language's inherent high polysemy. Initially, a native Arabic speaker skimmed the Arabic dictionary to identify polysemous terms. Subsequently, Wikipedia was surveyed to compile data and extract the most frequently utilized terms across various articles related but not limited to, art and culture, geography, health and fitness, history and events, mathematics and abstractions, natural sciences and nature, people and self, philosophy and thinking, religion, and spirituality. The native speaker then determined the most frequent terms, excluding stopwords. Finally, the top hundred frequent terms that carry polysemy based on the dictionary were chosen to build the WSD dataset.

A Python code, leveraging the Beautiful Soup library, was then employed to extract sentences from the Wikipedia Foundation containing these polysemous words. This automatic approach has resulted in a substantial amount of data. However, a subsequent manual sorting process was necessitated due to the absence of word meanings in Wikipedia. After collecting data, a native Arabic speaker sorted the data from the crawler based on polysemous target words. Word

Table 4
Disambiguation results.

Model	F1 score
KNN	0.71
Enhanced BERT [2]	0.96

senses were labeled based on text, keeping only diverse sentences to eliminate redundancy and repetition. This cleaning step and manual sorting have resulted in 155 examples taken from Wikipedia, as shown in Table 3.

Further efforts were made to enhance the dataset by conducting additional internet searches to acquire ten example sentences for each sense. The aim was to ensure the inclusion of diverse and comprehensive data from various reputable websites. Despite thorough research and extraction, certain word senses lacked examples on these websites, prompting further web searches. Additionally, a frequently used polysemous term might carry a sense where it is rarely used in context. For example, the term "لدج" (jdl), which is widely used in Arabic, especially as the meaning "شاقنلا" (nqA$) meaning debate, is rarely used as its second meaning, "يوَق" (qawiya) meaning became stronger. The most frequently used word, "لدج" (jdl), is rarely used in the context of "يوَق" (qawiya). Thus, a few examples of this term were found on the web. For this reason, GPT3.5-turbo [20] was used to generate sentences containing the word "لدج" (jdl) with "يوَق" (qawiya) context.

The data creation process using GPT3.5-turbo involved multiple steps. For less common senses, where internet-acquired data was available, GPT3.5-turbo was trained on the word, an example sentence from the web, and its sense to generate Arabic sentences. However, for words with limited or no efficient data found on the web, GPT3.5 was fed with the polysemous word, its sense, and a sentence example written by a native Arabic speaker. This approach was done since large language models are limited when dealing with rare words, necessitating fine-tuning for optimal performance.

In addition to human evaluation by a native Arabic speaker for the synthetic samples generated by GPT3.5-turbo, a classification model was employed for further validation. The K-Nearest Neighbors (KNN) algorithm was applied to disambiguate the data, ensuring robustness and reliability. The dataset was split, with synthetic samples reserved only for training, while testing involved only real samples. The result of this disambiguation method is presented in Table 4, showcasing the efficacy and utility of the classification model. Moreover, the original research paper [2] underscores the validity of the dataset for the WSD task by introducing an enhanced BERT model and comparing its performance with existing literature. The F1 score reported by the original paper [2] is also presented in Table 4. The table shows that both KNN and the enhanced BERT model are capable of disambiguating Arabic polysemous terms, achieving F1 scores of 0.71 and 0.96, respectively. This high performance gives evidence of the absence of bias in predicting the correct sense of a given word due to the presence of synthetic data. However, the superiority of the enhanced BERT model over KNN can be attributed to its contextualized nature due to the extensive training on a vast amount of textual data. Consequently, it exhibits heightened adeptness in adapting to the WSD task, thereby discerning context and semantics.

A case study was presented by testing the effect of WSD model trained on a sentiment analysis task as outlined in the original article [2] to assess this dataset's effectiveness and potential applications in Arabic NLP tasks. Given that Arabic NLP research lags behind that of other languages, WSD becomes crucial for enhancing the efficacy of implemented algorithms. Automated algorithms misinterpreting polysemous words or representing them similarly, even when they occur in varying contexts, can result in a high false positive rate. In the original article [2], it was demonstrated that even large language models still encounter challenges in downstream tasks such as machine translation, illustrated by the translation of the word "ابا" in the context of "church father/priest" as "father." This underscores the significance of establishing WSD datasets and models to bolster performance in downstream tasks. The case study focusing on sentiment analysis compared the model's performance before and after disambiguating polyse-

Table 5
Case study, sentiment analysis results as reported in [2].

Model	Precision	Recall	F1 score
Applying WSD prior sentiment analysis	0.9406	0.7070	0.8072
Applying sentiment analysis directly	0.9237	0.6812	0.7841

mous words, revealing that utilizing this data for constructing a WSD model led to more effective sentiment analysis. The results detailed in the original article [2] are presented in Table 5. The table demonstrates that applying disambiguation prior to the sentiment analysis task increased precision, recall, and F1 score by 1.69 %, 2.58 %, and 2.31 %, respectively.

Limitations

While this dataset offers valuable insights with diverse and comprehensive examples of various polysemous word senses in Arabic, it has some limitations. The dataset is limited to a hundred Arabic polysemous words, which could constrain a word-based disambiguation model to only these specific senses. Utilizing the data in its current form as supervised data for classification would not permit the implementation of a model capable of disambiguating words other than those appearing in the dataset. Thus, leading to poor performance in unseen or less frequent polysemous words. However, employing a model that utilizes sequence classification based on contextual cues and the senses provided could help disambiguate other terms and enhance the generalization of the data. The synthetic data augmented to the WSD dataset increases its diversity. However, the effectiveness of synthetic data depends on the accuracy and relevance of the generated examples. A native Arabic speaker reviewed each example generated by GPT3.5-turbo to ensure its accuracy, efficiency, and consistency, avoiding the presence of bias. In the future, this data could be expanded to cover more polysemous terms. Increasing its size is anticipated to yield more robust results. This expansion aims to enhance the dataset for more effective and widespread applications.

Ethics Statement

Ethical handling of the data is the highest priority feature of our research activities. The information collected is solely used for research purposes; no commercial gains are extracted from it. Scraper and automation were applied only to Wikipedia, where content sharing is welcome. Through the collection, we have followed with transparency and conformance with community guidelines. We do not copy, reproduce, download, post, store, distribute, transmit, broadcast full articles published; otherwise, we have benefited from the context to support our research. The careful review of a native Arabic speaker guarantees linguistic fidelity and cultural sensitivity in our data. Moreover, our data collection meets the guidelines of the public platforms' accessibility to prevent any identifiable information or entities from being targeted. Every piece of data in our research is handled with the utmost care and caution or put in a context that allows it to be different from previous sources to avoid copying so our research would stay reliable. Synthetic data produced through GPT3.5-turbo are run through plagiarism checkers and subject to human expert analysis to ensure their originality and resilience. Ethical data practices form the basis of our integrity and trust in responsible research. This data does not include any identity or targeting specific entities. Finally, the authors collect the data, and it is not a duplicate of information sourced from any other location.

Data Availability

Dataset for Arabic Word Sense Disambiguation (Original data) (Mendeley Data).

CRediT Author Statement

Sanaa Kaddoura: Data curation, Conceptualization, Methodology, Validation, Formal analysis, Writing – review & editing, Project administration, Funding acquisition; **Reem Nassar:** Software, Validation, Visualization, Data curation, Methodology, Formal analysis, Writing – original draft.

Acknowledgments

This work is funded by Zayed University Research Incentive Fund (RIF) with grant number [R22047].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Kaddoura, R. D. Ahmed, A comprehensive review on Arabic word sense disambiguation for natural language processing applications, Wiley Interdiscipl. Rev.: Data Mini. Knowl. Discov. 12 (4) (2022) e1447.
[2] S. Kaddoura, R. Nassar, EnhancedBERT: a feature-rich ensemble model for arabic word sense disambiguation with statistical analysis and optimized data collection, J. King Saud Univ.-Comput. Inf. Sci. 36 (1) (2024) 101911.
[3] S. Kaddoura, S.A. Alex, M. Itani, S. Henno, A. AlNashash, D.J. Hemanth, Arabic spam tweets classification using deep learning, Neural Comput. Appl. (2023) 1–14.
[4] J. Zakraoui, M. Saleh, S. Al-Maadeed, J.M Alja'am, Arabic machine translation: a survey with challenges and future directions, IEEE Access 9 (2021) 161445–161468.
[5] M. El-Razzaz, M. Fakhr, F. Maghraby, Arabic gloss WSD using BERT, Appl. Sci. 11 (2021) 2567, doi:10.3390/app11062567.
[6] A. Farghaly, A. Farghaly, K. Shaalan, Khaled, Arabic natural language processing: challenges and solutions, ACM Trans. Asian Lang. Inf. Process. (TALIP) 8 (4) (2009).
[7] Sanaa Kaddoura, Reem Nassar, Dataset for Arabic word sense disambiguation, Mendeley Data V1 (2024), doi:10.17632/pmdbs9tby8.1.
[8] Foundation, W., 2023. Wikipedia the free encyclopedia. https://ar.wikipedia.org/wiki/.
[9] AlJazeera, 2023. Aljazeera net: latest news of the day from around the world. Al Jazeera Media Network. https://www.aljazeera.net/.
[10] AlJazeera, 2023, Aljazeera documentary: stay updated on what's happening around the world. Al Jazeera Media Network. https://doc.aljazeera.net/.
[11] Altibbi, 2023. Altibbi website for health information and medical consultations: diseases, medications, and treatment. Altibbi FZ-LLC. https://altibbi.com/.
[12] Arabiaweather: Weather news & Forecast for Today and Tomorrow, ArabiaWeather, Inc, 2023 https://www.arabiaweather.com/.
[13] Shifa, 2023. Shifaa: a 24/7 renewed medical platform. shifaa platform. https://www.shifaa.ma/.
[14] CNN, 2023. CNN Arabic - latest political, sports, and entertainment news and video reports. Cable News Network. https://arabic.cnn.com/.
[15] BBC, 2023. BBC News Arabic - homepage. https://www.bbc.com/arabic.
[16] ArabicPost, 2023. ArabicPost. Integral Media Danismanlik Şti Limited or its licensors. https://arabicpost.net/.
[17] Argaam, 2023. Argaam: news and information about the Saudi stock market - tadawul. Argaam Investment. https://www.argaam.com/.
[18] Almaany, 2023. Multilingual and multidisciplinary dictionary of meanings - arabic-arabic dictionary. https://www.almaany.com/.
[19] King Salman International Complex for Arabic Language, 2023. Riyadh Dictionary of Contemporary Arabic Language. https://dictionary.ksaa.gov.sa/.
[20] OpenAI, 2023. ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat