Hindawi
Complexity
Volume 2020, Article ID 8896659, 7 pages
https://doi.org/10.1155/2020/8896659

Research Article

A Silver Standard Biomedical Corpus for Arabic Language

Nada Boudjellal,[1] **Huaping Zhang**,[1] **Asif Khan**,[1] **Arshad Ahmad**,[2]
Rashid Naseem,[3] **and Lin Dai**[1]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[2]*Department of Computer Science, City University of Science and Information Technology, Peshawar, Pakistan*
[3]*Department of IT and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences & Technology, Haripur, Pakistan*

Correspondence should be addressed to Huaping Zhang; kevinzhang@bit.edu.cn

Received 5 August 2020; Revised 21 September 2020; Accepted 25 September 2020; Published 9 October 2020

Academic Editor: Shafiq Ahmad

The rapidly growing data in many areas, as well as in the biomedical domain, require the assistance of information extraction systems to acquire the much needed knowledge about specific entities such as proteins, drugs, or diseases practically within a short time. Annotated corpora serve the purpose of facilitating the process of building NLP systems. While colossal work has been done in this area for English language, other languages like Arabic seem to lack these resources, especially in the healthcare area. Therefore, in this work, we present a method to develop a silver standard medical corpus for the Arabic language with a dictionary as a minimal supervision tool. The corpus contains 49,856 sentences tagged with 13 entity types corresponding to a subset of UMLS (Unified Medical Language System) concept types. The evaluation of a subset of corpus showed the efficiency of the method used to annotate it with 90% accuracy.

1. Introduction

With the exponential growth of data in many areas (news and economics), the task of processing the data and extracting useful information from it becomes a necessity. The biomedical domain is no exception. With more than 30 million citations of biomedical literature found in PubMed and an endless amount of electronic health records (EHRs), it is hard for researchers and practitioners of the domain to grasp the massive flow of data and get the needed knowledge from it in a practical way within a short time. Therefore, information extraction (IE) systems, in which natural language processing (NLP) techniques are used to turn the unstructured text data to an easily readable, well-structured text [1], are needed.

Corpora, which can be defined as a set of machine-readable texts sampled to represent a particular language or a variety of languages [2], have an essential role in NLP research since they provide a linguistic resource to build and test NLP systems.

Annotated corpora, which have been enriched with additional information, related to either structures (i.e., documents and paragraphs) or individual tokens such as part of speech (POS) tags or named entity identification, play an essential role in this task.

Since a considerable portion of biomedical literature is in English language, a significant part of NLP systems and existing corpora were dedicated to this language, where other languages like Arabic reveal a gap in both NLP systems and linguistic resources for the biomedical domain.

The classical manual methods of labeling data such as rule-based or supervised methods are known to be costly and time- and effort-consuming [3], that is why the idea of using silver corpora came to light; and it proved according to [4] that it could be effectively able to replace golden corpora for the task of NER. Besides, silver corpora can be generated with bigger sizes than golden corpora.

This paper presents a method to build an annotated biomedical Arabic corpus. Unlike other corpora, which were mostly manually labeled, the present corpus used a built bilingual dictionary for automatic annotation that does not need human intervention. Thus, the process of building the corpus comprises three main steps: acquisition of corpus documents, dictionary construction, and automatic corpus annotation. In order to assess the quality of the presented work, an evaluation has been performed on a sample of the annotated corpus.

This work has the following main contributions:

(i) The dictionary itself without a corpus can serve as a general medical linguistic resource that can be used to learn an Arabic Named Entity (NE) tagger. This method proved to be useful [5].

(ii) This method uses minimal supervision to annotate the corpus and thus reduce cost, time, and human effort.

(iii) The corpus can be used to test the efficiency of automatically annotated corpora in NLP systems or to train and test NER systems since it is annotated with 13 different entity types.

(iv) The corpus is a linguistic resource for a language other than English language.

The importance of this work can be resumed in some key points stated as follows:

(i) The dictionary can be used as seed start to train a minimally supervised classifier for an enhanced annotation of a medical Arabic or English or bilingual corpus or to enrich a general purpose corpus with medical annotations.

(ii) The corpus can be used to test the effectiveness of silver corpora in NLP tasks,

(iii) The corpus can be optimised to suit a specific task-like disease classification or for relation extraction task such as disease-treatment relations for a distant supervised setting, and the work of [6] is an example of such practices.

(iv) The corpus can be applied to state-of-the-art deep learning systems such as BERT [7] for NER task.

(v) Overall, this work can be seen as a boost for linguistic biomedical research for Arabic language.

The remaining of the paper is organized as follows: related work concerning corpora in languages other than English plus existing silver corpora are reviewed in Section 2. In Section 3 the process of dictionary and corpus construction is described. The method of corpus annotation is given in Section 4. Section 5 describes corpus evaluation and results. Finally, Section 6 concludes the paper and outlines some directions for future work.

2. Related Work

Although a lot of work has been done for English corpora, the work on other languages is still growing. In this section, a brief review is presented about corpora in languages other than English and some of the existing silver corpora.

For the Romanian language, the MoNERo corpus [8] was created as a medical gold standard corpus with morphological and named entities annotations. The corpus consists of 4,989 sentences from articles related to cardiology, diabetes, and endocrinology and was annotated with four NE types: anatomy, chemicals and drugs, disorders, and procedures.

Quaero corpus [9] was built for the French Language. It contains 103,056 words collected from three types of documents, where the authors used ten entity types corresponding to UMLS [10] semantic groups. They used an automatic annotation, which was validated later by human experts.

The Swedish language has its part also with a semantically tagged corpus [11] that contains the electronic versions of the Journal of the Swedish Medical Association. After developing the corpus, it was used for term validation and term extraction tasks.

For the Arabic language, a corpus for drug information was built [12]. The corpus contains documents about 202 drugs, and it was manually annotated with four entity types, including drug generic name, its brand, its chemical formula, and class.

These corpora and others mentioned in [13] are golden (i.e., manually annotated) that serve as useful linguistic resources for NLP systems in languages other than English. Still, as any manually conducted task, it consumes a lot of time and effort. One alternative way to save them is the creation of silver clinical corpora. Existing silver corpora are scarce, and they can be categorised in 3 categories: (i) silver corpora with NE annotations: an example for this category is the CALBC silver standard corpus [14] that was built by combining the output of different automatic annotation methods (rule-based and dictionary-based), to get a sole-coordinated NE annotated corpus with different biomedical semantic types (e.g., disease, protein, chemical, and drugs); (ii) silver corpora with relation annotations: the corpus of [15] falls in this category.

NER tools were used to annotate human phenotype and gene entities—with 87% of precision—from abstracts from PubMed. Then, by using a distant supervision approach, the authors classified human phenotype-gene relations using the HPO file that contains gold-standard human phenotype-gene relations. Then, they evaluated the corpus by using 2 deep learning approaches: BO-LSTM (precision = 69.23%) and BioBERT (precision = 78.95%). (iii) Silver corpora with syntaxic annotations: the authors of [16] created an Arabic language word segmented corpus composed of 18,167,183 words from newspaper articles. A rule-based approach was used for segmentation after POS tagging which was done with Stanford POS Tagger.

The presented work describes the process of creating a silver standard biomedical corpus for the Arabic language with NE through a dictionary-based method.

3. Corpus Construction

This section describes the process of corpus development, which includes three main steps mentioned and depicted in

Figure 1, from the selection of corpus documents and text preprocessing and then dictionary construction to the corpus annotation.

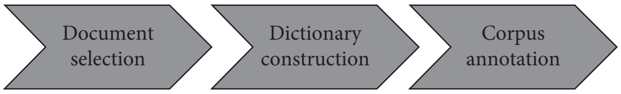


Figure 1: Process of corpus development.

3.1. Selection of Corpus Documents and Preprocessing. The documents used to build the present corpus were obtained from altibbi.com [18], which is a well-managed Arabic medical website containing a free English-Arabic medical dictionary with terms description and a considerable number of healthcare-related articles. The dictionary provided by this resource will be used in the next step. For the time being, the corpus contains 49,856 sentences distributed into 978 documents related to different medical topics (for example, cardiology, pulmonology, dermatology, gynecology, and traditional medicine).

3.1.1. Tokenization and Part of Speech Tagging. The corpus has two versions (Table 1):

(1) The first version is raw text annotated with NE, where the dictionary was used to tag the corpus. Table 1 gives briefly some general statistics of the corpus.

(2) The second version of the corpus contains a tokenized and tagged text with POS. The Stanford Arabic Word Segmenter [19] was used for the tokenization and Stanford Arabic Part of Speech Tagger [20] for the POS tagging task. Then, the IOB2 encoding format was used for NE tagging (see Section 3.3).

Table 1: General statistics of the corpus.

Tokens	1,195,805
Words	999,832
Sentences	49,856
Documents	978

3.2. Dictionary Building. Dictionaries proved to be a useful source of annotation or a tool to build NE taggers [5, 21]. Since linguistic resources related to the medical domain are very scarce in the Arabic language, such as knowledge bases or annotated corpora, using a dictionary for NE annotation seems like an appropriate choice.

3.2.1. Optimization of UMLS Data. UMLS (Unified Medical Language System) is a set of biomedical terminologies that are grouped under common concepts with their relations and the underlying terms within each concept, which allows an easy translation and interoperability. It comprises a set of 133 different concept types, from which 13 concept types were selected for this work, namely, Disease or Syndrome, Sign or Symptom, Therapeutic or Preventive Procedure, Diagnostic Procedure, Mental or Behavioral Dysfunction, Antibiotic, Virus, Hormone, Enzyme, Clinical Drug, Injury or Poisoning, Bacterium, and Gene or Genome.

The reason behind selecting these specific types was for the future use of the corpus, which can be for Named Entity Recognition (NER) or Relation Extraction (RE) tasks that both serve information extraction. An example of relations that can be extracted is drug-disease relations. The output of this step is a set of English language medical terms with their corresponding concept type.

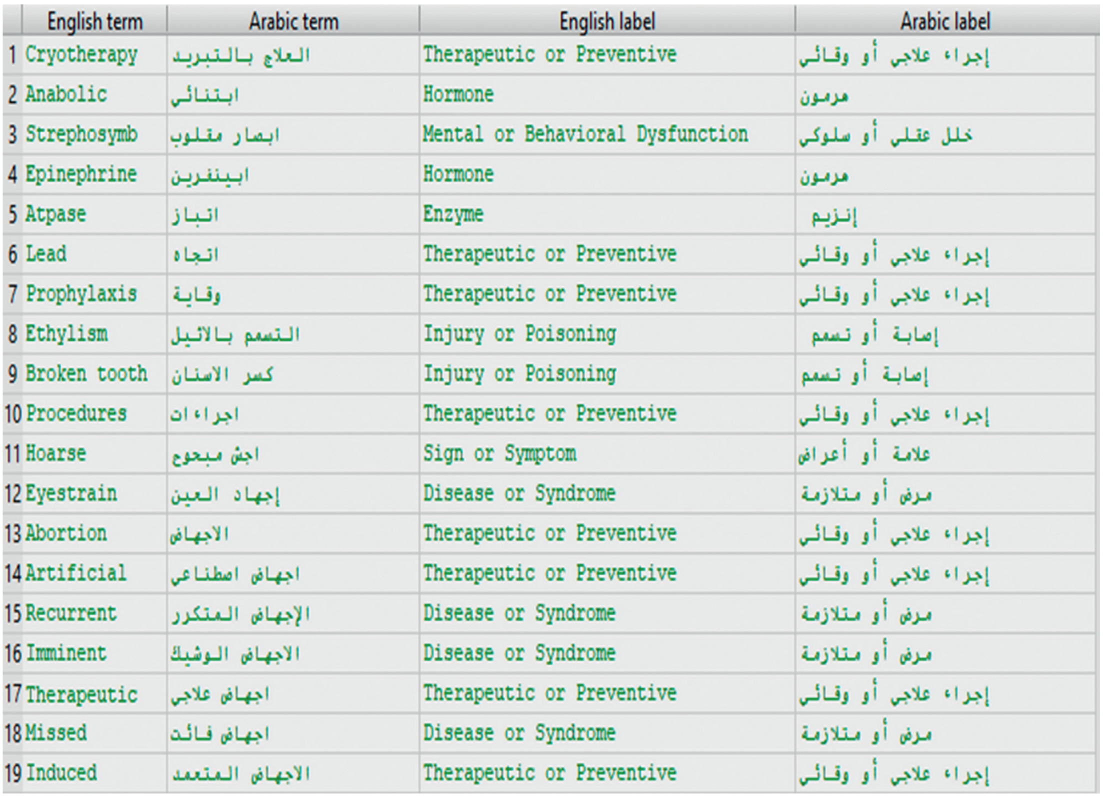
Table 2: The number of terms for each concept type in the annotated English-Arabic medical dictionary.

Concept type	Number of terms
Disease or syndrome	3,748
Therapeutic or preventive procedure	1,119
Sign or symptom	508
Mental or behavioral dysfunction	386
Diagnostic procedure	303
Injury or poisoning	265
Gene or genome	218
Enzyme	166
Bacterium	160
Virus	117
Hormone	71
Antibiotic	9
Clinical drug	4

3.2.2. Bilingual Medical Dictionary (English-Arabic). Because we believe that translating the English medical terms obtained in the first step is time-consuming and very prone to error since online dictionary translation is not that efficient, scrapping an existing medical dictionary was a better option. Thus, the online freely available bilingual medical dictionary provided by altibbi.com was used. The output of this step is a medical English-Arabic dictionary of 33,743 pairs.

3.2.3. Mapping and Disambiguation. To ensure that all data are accurate, to map the output of step 2 with that of step 1, only terms that have an exact match with UMLS terms were taken into consideration. Thus, the resulted dictionary has fewer pairs. To reduce the negative effect of ambiguity on the tagging process, ambiguous Arabic terms that can have different meanings or can be used as both a verb and a noun were manually removed from the dictionary; for example, the word "حالة" which corresponds to "Lysin" or "Enzyme" in the dictionary can have two meanings depending on whether it is written with "shadda," i.e., "حالّة" or "Lysin" or without "shadda," i.e., "حالة" or "Status". If this word is kept in the dictionary, every word "حالة" in the text will be labeled as "Enzyme," and it will create a lot of noise.

Table 2 describes the distribution of concept types in the dictionary by the number of terms for each concept type. Figure 2 gives a perspective of the dictionary.



	English term	Arabic term	English label	Arabic label
1	Cryotherapy	العلاج بالتبريد	Therapeutic or Preventive	إجراء علاجي أو وقائي
2	Anabolic	ابتنائي	Hormone	هرمون
3	Strephosymb	ابصار مقلوب	Mental or Behavioral Dysfunction	خلل عقلي أو سلوكي
4	Epinephrine	ابينفرين	Hormone	هرمون
5	Atpase	اتباز	Enzyme	إنزيم
6	Lead	اتجاه	Therapeutic or Preventive	إجراء علاجي أو وقائي
7	Prophylaxis	وقاية	Therapeutic or Preventive	إجراء علاجي أو وقائي
8	Ethylism	التسمم بالاثيل	Injury or Poisoning	إصابة أو تسمم
9	Broken tooth	كسر الاسنان	Injury or Poisoning	إصابة أو تسمم
10	Procedures	اجراء ات	Therapeutic or Preventive	إجراء علاجي أو وقائي
11	Hoarse	اجش مبحوح	Sign or Symptom	علامة أو أعراض
12	Eyestrain	إجهاد العين	Disease or Syndrome	مرض أو متلازمة
13	Abortion	الاجهاض	Therapeutic or Preventive	إجراء علاجي أو وقائي
14	Artificial	اجهاض اصطناعي	Therapeutic or Preventive	إجراء علاجي أو وقائي
15	Recurrent	الإجهاض المتكرر	Disease or Syndrome	مرض أو متلازمة
16	Imminent	الاجهاض الوشيك	Disease or Syndrome	مرض أو متلازمة
17	Therapeutic	اجهاض علاجي	Therapeutic or Preventive	إجراء علاجي أو وقائي
18	Missed	اجهاض فائت	Disease or Syndrome	مرض أو متلازمة
19	Induced	الاجهاض المتعمد	Therapeutic or Preventive	إجراء علاجي أو وقائي

Figure 2: A sample of annotated bilingual medical dictionary.

4. Corpus Annotation

This section describes the corpus annotation method.

4.1. Raw Text Annotation. The Arabic language has a complex morphological nature. For example, the clitics in the Arabic language are agglutinated to words, while in other languages, they are treated as single words. Its agglutinative nature is considered as a challenge in building Arabic NLP systems [22], alongside with the lack of capitalization (which helps in languages like English to identify most of NE) and short vowels (which are replaced with diacritics in Arabic language; these diacritics, when not used, can cause a disambiguation problem since there are words with the same characters, but when different diacritics are used, the meaning will be changed).

Considering all the above specifications of the Arabic language, a fuzzy matching algorithm with n-grams was used to tag the NE in the raw text of the Arabic medical corpus. Other methods were not used simply because of the following:

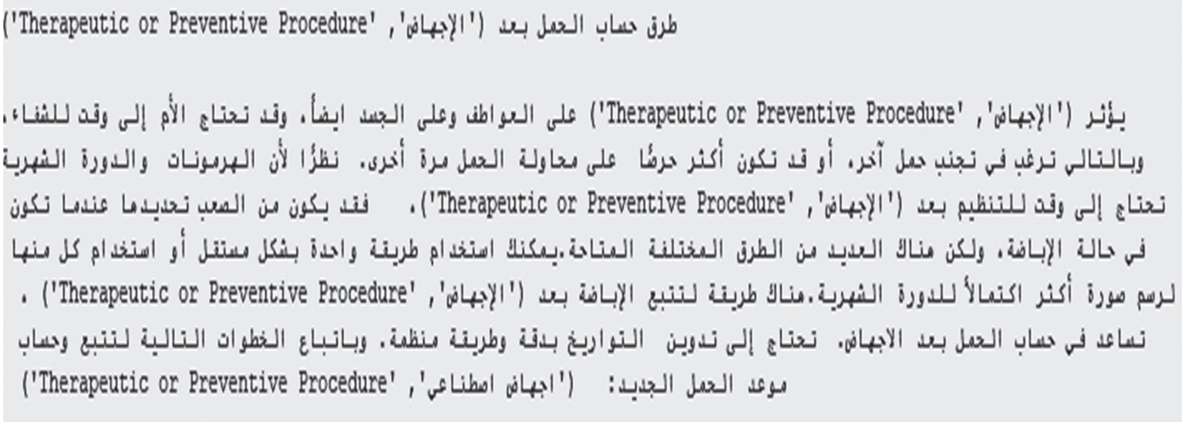
(i) High complexity such as deep learning methods with their incorporation in NLP tasks showed good performance mainly in NER and relation extraction, but this progress seems to only benefit languages with rich resources and huge labeled data (mainly English language), whereas the scarcity or nonexistence of training data especially clinical one presents a limitation and challenge for DL methods because a huge amount of data is needed to process and get the desired results for supervised DL-based NER [23].

(ii) Being costly and time- and effort-consuming (namely, rule-based methods).

Fuzzy matching or approximate string matching refers to the process of identification of data items that are not the same but can vary up to prefixed limitations [22] Unlike exact matching, which gives only entities that have the same syntactic form of terms in the dictionary, fuzzy match allows the detection of entities that have suffixes, prefixes, or clitics agglutinating to them in the Arabic language.

In our algorithm, a threshold was set heuristically for similarity. When choosing a threshold less than 0.9 (for example, 0.85), the program output tends to have a lot of noisy data especially words and expressions that differ from dictionary entries in one letter or two apart from conjunction letters "و (waw) and ف (fa)," prepositions "ب (bi) and ل (li) and ك (ka)," and "ال التعريف (al altaerif)" which is the equivalent of "the" in English. These letters are always written attached to words in Arabic language.

On the contrary, if the threshold is more than 0.9, the program ignores a lot of valid words, and thus, the recall and accuracy draw down. Therefore, the threshold was fixed more or equal to 0.9, which gave the most accurate annotations. Figure 3 presents a sample from the annotated corpus. The named entities are put between brackets along with their entity type, i.e., (e, a), where "e" is a named entity and "a" is its semantic label, as it is shown in Figure 3. The



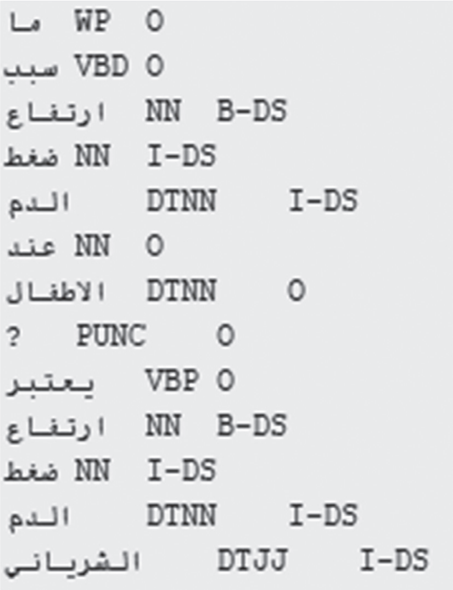
```
طرق حساب الحمل بعد ('الإجهاض', 'Therapeutic or Preventive Procedure')

يؤثر ('الإجهاض', 'Therapeutic or Preventive Procedure') على العواطف وعلى الجسد ايضاً، وقد تحتاج الأم إلى وقت للشفاء،
وبالتالي ترغب في تجنب حمل آخر، أو قد تكون أكثر حرصًا  على محاولة الحمل مرة أخرى. نظرًا لأن الهرمونات  والدورة الشهرية
تحتاج إلى وقت للتنظيم بعد ('الإجهاض', 'Therapeutic or Preventive Procedure')،    فقد يكون من الصعب تحديدها عندما تكون
في حالة الإباضة، ولكن هناك العديد من الطرق المختلفة المتاحة.يمكنك استخدام طريقة واحدة بشكل مستقل أو استخدام كل منها
لرسم صورة أكثر اكتمالاً للدورة الشهرية.هناك طريقة لتتبع الإباضة بعد ('الإجهاض', 'Therapeutic or Preventive Procedure') ،
تساعد في حساب الحمل بعد الاجهاض. تحتاج إلى تدوين  التواريخ بدقة وطريقة منظمة. وباتباع الخطوات التالية لتتبع وحساب
موعد الحمل الجديد:  ('اجهاض اصطناعي', 'Therapeutic or Preventive Procedure')
```

Figure 3: A sample of annotated text.

resulted corpus has 73,284 annotated named entities, including 7,601 labeled as "Disease or Syndrome," 3,166 as "Therapeutic or Preventive Procedure," and 1,583 as "Sign or Symptom."

4.2. IOB2 Annotated Corpus. To adapt the corpus to be used in further research purposes easily, IOB2 (inside-outside-beginning) encoding format was adopted to tag the tokenized version of the corpus based on the result of the first version in the insight of being widely used in most NLP related work. B-tag is used to indicate the beginning of a named entity, while I-tag and O-tag are used to indicate that the correspondent token is inside the named entity and outside of it, respectively. An example is shown in Figure 4. In this example, the NE is "ارتفاع ضغط الدم" (hypertension), and it is labeled as "Disease or Syndrome," which is referred to in the example as "DS."

The other semantic types are used as follows: in the IOB2 annotation, Therapeutic or Preventive Procedure (B-TPP and I-TPP), Sign or Symptom (B-SS and I-SS), Mental or Behavioral Dysfunction (B-MBD and I-MBD), Diagnostic Procedure (B-DP and I-DP), Injury or Poisoning (B-IP and I-IP), Gene or Genome (B-GENE and I-GENE), Enzyme (B-IZ and I-IZ), Bacterium (B-BACT and I-BACT), Virus (B-VIRUS and I-VIRUS), Hormone (B-HR and I-HR), Antibiotic (B-AB and I-AB), and Clinical Drug (B-CD and I-CD).



```
ما  WP  O
سبب VBD O
ارتفاع  NN  B-DS
ضغط NN  I-DS
الدم    DTNN    I-DS
عند NN  O
الاطفال DTNN    O
?   PUNC    O
يعتبر   VBP O
ارتفاع  NN  B-DS
ضغط NN  I-DS
الدم    DTNN    I-DS
الشرياني    DTJJ    I-DS
```

Figure 4: An example of annotated tokenized text.

Table 3: Results of the experiment.

Total identified NE	117
Total NEs	149
Correctly identified (true positives)	105
Incorrectly identified (false positives)	12
Missed NE (false negatives)	44

5. Corpus Evaluation and Results

To evaluate the silver corpus, 300 sentences were randomly selected and manually tested by a domain expert to check the accuracy of annotated named entities. The expert was asked to mark each pair (entity and annotation) with C if the annotation is correct and F if it is incorrect and mark the missed/untagged named entities in sentences with U. At the end, for each sentence, the correctly tagged NE (true positives), the incorrectly tagged NE (false positives), and missed or untagged NE (false negatives) were identified. The results are summarized in Table 3.

The results show that out of 149 existing named entities in the selected subset, 117 entities were identified from which almost 90% were correctly identified (i.e., 105 named entities). Although the selected subcorpus used for testing is relatively small in accordance with the total corpus, the result obtained from it can be considered as a proof of the effectiveness of the work.

The reasons behind the unidentification of NE or misidentification of others can be resumed as follows:

(i) The number of Named Entities that can be identified is limited to the number of terms included in the dictionary (i.e., 7074), which makes it almost impossible to identify terms that do not belong to the dictionary.

(ii) Some entities were not identified because they did not fall within the scope of the threshold set, due to having more morphological differences with the current terms in the dictionary.

(iii) The similarity factor in the algorithm is the leading cause for the misidentification of NE since some words are very similar and only differ from each other with just one letter. For example, the word "الإحماء" (warm-up) is mistaken for the word "الإغماء" (syncope), which is labeled as "Sign or Syndrome."

6. Conclusion and Future Work

In this work, we have presented a method to create a medical corpus for the Arabic language and annotate it with 13 different types of entities using minimal supervision without human intervention. The corpus has two varieties: a morphologically annotated version according to IOB2 standards and a raw text annotated version. The evaluation of the test set gave an accuracy of 90%. To the best of our knowledge, this is the first corpus of this type for the Arabic language especially for clinical domain. This corpus can be used for bioNLP tasks for the Arabic language, and it can be improved in many ways:

(i) Dictionary-wise: expanding the dictionary with extra entries from other available bilingual resources can improve the performance and reduce the number of false negatives.

(ii) Corpus-wise: the raw text version of the corpus is flexible and can be improved with up-to-date clinical articles and different biomedical text resources.

Data Availability

The corpus data used to support the findings of this study are available from the corresponding author upon request. The UMLS data used to support the findings of this study were supplied by "National Library of Medicine-National Institutes of Health" under license and so cannot be made freely available by authors.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research work was funded by the National Key Research and Development Project of China (Grant no. 2018YFC0832304) and National Science Foundation of China (Grant no. 61772075). The authors are thankful to them for their financial support.

References

[1] A. Téllez-Valero, M. Montes-y-Gómez, and L. Villaseñor-Pineda, "A machine learning approach to information extraction," *Computational Linguistics and Intelligent Text Processing*, Springer, Berlin, Germany, pp. 539–547, 2005.

[2] T. McEnery, R. Xiao, and Y. Tono, *Corpus-based Language Studies: An Advanced Resource Book*, Routledge, Abingdon, UK, 2006.

[3] N. Boudjellal, H. Zhang, A. Khan, and A. Ahmad, "Biomedical relation extraction using distant supervision," *Scientific Programming*, vol. 2020, Article ID 8893749, 9 pages, 2020.

[4] N. Kang, E. M. van Mulligen, and J. A. Kors, "Training text chunkers on a silver standard corpus: can silver replace gold?" *BMC Bioinformatics*, vol. 13, no. 1, p. 17, 2012.

[5] J. Shang, L. Liu, X. Ren, X. Gu, T. Ren, and J. Han, "Learning named entity tagger using domain-specific dictionary," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 2054–2064, Brussels, Belgium, October 2018.

[6] K. E. Ravikumar, H. Liu, J. D. Cohn, M. E. Wall, and K. Verspoor, "Literature mining of protein-residue associations with graph rules learned through distant supervision," *Journal of Biomedical Semantics*, vol. 3, no. 3, 2012.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, MN, USA, October 2018.

[8] M. Mitrofan, V. B. Mititelu, and G. Mitrofan, "MoNERo: a biomedical gold standard corpus for the Romanian language," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 71–79, Florence, Italy, August 2019.

[9] A. Névéol, C. Grouin, J. Leixa, S. Rosset, and P. Zweigenbaum, "The *Quaero* French medical corpus: a ressource for medical entity recognition and normalization," in *Proceedings of the BIOTEXTM*, Reykjavik, Iceland, 2014.

[10] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, 2004.

[11] D. Kokkinakis and U. Gerdin, "A Swedish scientific medical corpus for terminology management and linguistic exploration," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010.

[12] H. Al-Ibrahim, H. S. Al-Khalifa, and A. M. Al-Salman, "Towards building Arabic corpus for drug information," in *Proceedings of the MEDES 2014—6th International Conference on Management of Emergent Digital EcoSystems*, pp. 67–71, Buraidah Al Qassim, Saudi Arabia, June 2014.

[13] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, and P. Zweigenbaum, "Clinical natural language processing in languages other than english: opportunities and challenges," *Journal of Biomedical Semantics*, vol. 9, no. 1, pp. 1–13, 2018.

[14] D. Rebholz-Schuhmann, A. J. J. Yepes, E. M. Van Mulligen et al., "CALBC silver standard corpus," *Journal of Bioinformatics and Computational Biology*, vol. 8, no. 1, pp. 163–179, 2010.

[15] D. Sousa, A. Lamurias, and F. M. Couto, "A silver standard corpus of human phenotype-gene relations," in *Proceedings of the 2019 Conference of the North*, pp. 1487–1492, Minneapolis, MN, USA, June 2019.

[16] H. Awdeh, A. Abdallah, G. Bernard, M. Hajjar, and M. El-Sayed, *A Silver Standard Arabic Corpus for Segmentation and Validation*Eliva Press, Chisinau, Republic of Moldova, 2019.

[17] "Altibbi site for health information and medical advice, diseases, drugs and cures," 2020, https://www.altibbi.com.

[18] W. Monroe, S. Green, and C. D. Manning, "Word segmentation of informal Arabic with domain adaptation," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 206–211, Baltimore, MA, USA, June 2014.

[19] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the NAALC*, Edmonton, Canada, May 2003.

[20] S. N. Kim and L. Cavedon, "Classifying domain-specific terms using a dictionar," *Australasian Language Technology Association Work. (ALTA 2011)*, pp. 57–65, 2011.

[21] S. Alanazi, B. Sharp, and C. Stanier, "A named entity recognition system applied to Arabic text in the medical domain," *International Journal of Computer Science and Information*, vol. 12, no. 3, 2015.

[22] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020, In press.

[23] T. Rees, "Taxamatch, an algorithm for near ("Fuzzy") matching of scientific names in taxonomic databases," *PLoS One*, vol. 9, no. 9, Article ID e107510, 2014.