Medical Term Extraction in an Arabic Medical Corpus

Doaa Samy♦ *, Antonio Moreno-Sandoval*, Conchi Bueno-Díaz*, Marta Garrote-Salazar* and José M. Guirao□

♦ Cairo University, Faculty of Arts, Egypt
*Computational Linguistics Laboratory-Autónoma University Madrid
□Granada University

doaasamy@cu.edu.eg, antonio.msandoval@uam.es, diazmunio@hotmail.com, marta.garrote@uam.es, jmguirao@ugr.es

Abstract

This paper tests two different strategies for medical term extraction in an Arabic Medical Corpus. The experiments and the corpus are developed within the framework of Multimedica project funded by the Spanish Ministry of Science and Innovation and aiming at developing multilingual resources and tools for processing of newswire texts in the Health domain. The first experiment uses a fixed list of medical terms, the second experiment uses a list of Arabic equivalents of very limited list of common Latin prefix and suffix used in medical terms. Results show that using equivalents of Latin suffix and prefix outperforms the fixed list. The paper starts with an introduction, followed by a description of the state-of-art in the field of Arabic Medical Language Resources (LRs). The third section describes the corpus and its characteristics. The fourth and the fifth sections explain the lists used and the results of the experiments carried out on a sub-corpus for evaluation. The last section analyzes the results outlining the conclusions and future work.

Keywords: Arabic Medical Language Resources, Arabic Medical Terms, Term Extraction.

1. Introduction

This paper presents an experiment carried out within MULTIMEDICA project. The experiment goal is to test two different strategies for medical term extraction in an Arabic corpus: the first one is based on a list of specific medical terms in Arabic in their full form; and the second one is a list of Arabic equivalents of Latin prefix and suffix commonly used in the medical and health domain. Arabic equivalents are words that can form part of compound terms.

For example, the first list includes terms in its complete form such as the term "conjunctivitis" and its Arabic translation "التهاب الملتحمة". The second list includes only the Latin suffix "-itis" and its Arabic equivalent which is in this case is "التهاب".

As a test dataset, an Arabic Medical corpus has been built from Health sections in Arabic newswire texts and health portals. Thus, the experiments carried out and described in this paper offer the community new resources in the Arabic medical and health domain (corpus and terminological database).

Multimedica is a project funded by the Spanish Ministry of Science and Innovation. The project aims at developing multilingual resources and tools for processing of newswire texts in the Health domain. Languages covered in the project are: Spanish, Arabic and Japanese. Developed resources and tools will be included in a translation and terminology portal targeting students and professors at Spanish universities. This portal will include a term extractor applied to comparable corpora in Spanish, Arabic and Japanese.

In this paper, we will outline the methodology applied on Arabic language. The abstract is divided into four sections: a review of the state-of-art in Arabic medical Language Resources (LR), building the corpus, terminological lists and, finally, experiments and results.

2. State-of-the-Art in Arabic Medical LR

State-of-art in Arabic medical and health domain represents some challenges when addressing language resources and tools. These challenges are due to certain practices adopted by practitioners and specialists within the medical and health domain in many countries across the Arab World.

The main challenge in addressing Arabic Language Resources (LR) in health sciences is the clear diglossia phenomenon prominent among specialists and practitioners in the field. The basic definition of diglossia, according to Charles Ferguson (1959), refers to a linguistic phenomenon, mainly, a sociolonguistic phenomenon where two languages or two dialects are used by the same community in different social situations for different social purposes. Diglossia can be observed in the following aspects:

- First, Arabic is not the language used in teaching Medicine, Pharmacy and other health related programmes at the university level in many Arab countries. Instead, English or French are used as lingua franca. In Morocco, Tunisia and Algeria French is used, while in Egypt, Iraq, Jordan, Saudi Arabic and Gulf countries, English language is used. Syria is the only exception where Arabic is used in teaching and health practices.
- Second, English or French are the languages used in professional practices within the health domain in Arab

countries where documents, prescriptions, reports, etc. are mostly produced in one of these foreign languages.
- Third, scientific publications and literature in the health domain in Arab countries are not an exception. Scientific articles are mostly written either in English or French.
All the diglossic aspects are challenges to the proposed study, especially in achieving the test datasets and evaluating the experiments against this data. For example, a query for articles published in Arabic in the famous Medline/PubMed retrieved only one result. Specialized textbooks in health related subjects were not available.
The study of the state-of-art concerning computational approaches for language processing or language resources in the medical domain in Arabic revealed a complete gap in this area. Also Arabic lacks versions of resources such as UMLS or SNOMED. This might be due to the significant lack of Arabic textual resources and corpora in the health domain. To our knowledge, previous studies have not addressed this domain in Arabic language from a computational perspective. Few experiments on enhancing web browsing and searching in underdeveloped web were applied on medical Arabic web portals (Chung & Chen, 2009). Arabic medical web portals were chosen as an example of "underdeveloped web" defined by authors as "lack of high-quality content and functionality. An example is the Arabic Web, in which a lack of well-structured Web directories limits users' ability to browse for Arabic resources". However, these studies do not address linguistic features, LR or NLP issues.
On the other hand, the fact that Arab specialists in this domain can easily access the information and resources in other languages (English and French) does not represent a serious need to develop resources in Arabic. However, this should not be a pretext to abandon this domain because if we change the perspective take into consideration the patients or other key players in health services such as administrative or non-specialists staff, we could easily notice the need for developing such resources.
Patients and non-specialists face difficulties in communication due to the diglossic situation. They hardly understand the specialists' reports or their language. This difficulty in accessing and understanding information by non-specialists requires more resources and tools to help overcoming this linguistic barrier. Thus, tools and resources in Arabic are needed not only for translators and terminologists as a step towards a better information flow. Also, these resources/tools could play an important role in providing better health services and in guaranteeing patient's safety.
To bridge the communication gap between the specialists and the public, some newspapers have sections for health-related topics in which specialized information is simplified or adapted to reach the public and answer their inquiries. Also, medical portals, such as Altibbi, have been developed to provide some interaction between public and specialists in health related topics.
It is also important to point out the efforts carried out by initiatives aiming at Arabization in the medical and health domain. AHSN (Arabization of Health Sciences Network) is an initiative by the East Mediterranean Regional Office of the World Health Organization1. The Unified Medical Dictionary resulted from these efforts. It is a multilingual dictionary of medical terms including English, French, Spanish, German and Arabic.
Regarding arabization initiatives, it is necessary to highlight that most of the steps have been taken by Syrian and Iraqi specialists. This is why the Levantine Arabic variety is majorly used. This also represents another challenge since terminological variations are sometimes not easily understood by other Arab speaking countries.

3. Building an Arabic Medical Corpus

Given the lack of specialized sources in Arabic, we opted for newswire texts and medical portals. Regarding the text typology, we are aware that the type of texts available is not highly specialized since it is simplified to address the general public. However, it is considered as a feasible and valid option as it represents an intermediate linguistic register combining features of the specialized language together with the common linguistic features. On the other hand, it is a first step in this area that can be extended in future studies.
To build the corpus, we used texts available on Internet from the following sources2:

- Altibbi portal http://www.altibbi.com/. It is an online Arabic medical and health resource. The portal is a Jordanian initiative to provide the Arab community with health portals as the American WebMD or Healthline. As per the definition provided by Altibbi "The portal provides a medical dictionary, medical articles and news, as well as question and answer features [...]"
- Newswire medical text. In this respect, we collected texts from online newspapers which included a section for Health. The texts were automatically collected by retrieving documents from Health directories in the selected newspapers. The corpus includes three subcopora from three different newspapers representing three different geographical areas within the Arab World:
    - Asharq Al-Awsat (Middle East)- Saudi Arabia http://www.aawsat.com/
    - Youm7 (The Seventh Day)- Egypt http://www.youm7.com/
    - El Khabar (The News)- Algeria http://www.elkhabar.com

Geographic distribution was meant to observe if varieties of the Arabic language used affect the results. Although Modern Standard Arabic is the variety used in all, but there are still some features that could be characteristic of

[1] http://www.emro.who.int/ahsn/AboutAHSN.htm
[2] The use of these texts is for educational and research purposes and thus does not violate their copyrights.

certain regions. The following table shows the four subcorpora, the number of tokens and documents retrieved.

Source	Number of documents	Number of tokens
Altibbi	43278	2 398 876
Aawsat	68	48 493
Youm7	83	18 948
ElKhabar	97	21 032

Table 1: Arabic Medical Corpus (Documents and Tokens).

Text crawling procedure: we used the unix command "wget" to capture the documents. After analyzing the structure of each newspaper, an xml version of every document was generated, using unified metadata.

4. Elaboration of Arabic Medical Term Lists

The main idea of the experiment is to explore what is more useful in medical term extraction from a corpus of newspapers and a health portal. Two approaches were evaluated:

a. A general list of full terms in Arabic, generated from an English medical term list of 3473 entries extracted from resources such as SNOMED and UMLS. The list includes both single terms as well as multiword terms.

Terms were automatically translated into Arabic using Google translator. Then, the translated terms were validated by a native linguist using Arabic Wikipedia and two online (English-Arabic) dictionaries:

- Al-tibbi dictionary
- Unified Medical Dictionary provided by EMRO-WHO.

In case of multi-word terms and while validation, translating focused on nominal heads of the terms, while modifiers (e.g. adjectives) were only translated if they represent specialized terms and not general domain adjectives such as "extended", "low", etc.. For example, modifiers or adjectives such as "acute" were not translated each time they appeared. This validated list of 3473 terms is an accurate resource, but it is still an incomplete resource that needs to be further extended.

A possible way to extend the list is to use it as a list of seed terms. Each term is looked up in the dictionaries. The results retrieved included all possible combination of the seed term. These possible combinations are used to increase the initial seed list. However, for these experiments we did not use the extended list. We only experimented with the initial 3473 terms.

b. An elaborated list of English prefixes and suffixes (460) used in medical and health terms (eg. cardio-, -itis). The list of English prefixes and suffixes were automatically translated into Arabic (التهاب، قلب), then manually validated. The total number in Arabic was 410 since some prefixes and suffixes were too general and difficult to translate such as "re-" or "-tic, "-ous".

We opted for translating the complete meaning rather than the prefix or suffix because Arabic applies a different approach to create neologisms. While English or Spanish use derivation through Latin prefix and suffix, Arabic uses lexical composition for medical terms. Thus, Arabic uses the whole lexeme (inflammation) but not the derivational morpheme (-itis).

Although this reduced list of suffix/prefix might represent a less accurate approach, since the words can be part of a non-term candidates, but on the contrary it can produce better recall in finding new terms.

5. Experiments on Term Extraction in the Arabic Corpus

Two experiments were carried out using the different sub-copora.

In the first, the extended accurate list of complete terms was used. At this stage we used only the seed terms (3473 terms), while in the second the reduced list of possible "term components" equivalent to suffix and prefix (410) was used. In each experiment we tested how many terms are retrieved in the corpora. Table 2 shows results for the list A (full terms).

Subcorpus	Term types identified	Total term list	Term Tokens identified-ocurrence
Al-tibbi	919	3473	73 706
Aawsat	184	3473	701
Youm7	100	3473	327
El-Khabar	126	3473	502

Table 2: Results for Term list A on the whole corpus

Subcorpus	Term types identified	Total term list	Term Tokens identified-ocurrence
Al-tibbi	404	410	338 637
Aawsat	317	410	3 359
Youm7	258	410	2 078
El-Khabar	265	410	2 467

Table 3. Results for the Affix list B (basic compositional terms) on the whole corpus

The above tables show that using the reduced list-B with basic compositional terms give better results. It is more feasible and represents an efficient yet quick approach since it is less time consuming and less laborious task.

To evaluate the above experiments, we used a sample dataset of the different sub-corpora formed up from a total of 2273 tokens.

Evaluation was carried out in different rounds. In every experiment, the test dataset was looked up for occurrences from the term list**.** The identified terms were tagged in the test dataset. In this respect, it is important to highlight that the extracted terms are to be annotated in the corpus and, thus, providing an integrated language resource in the Arabic medical domain.

Experiment1-Evaluation-Round 1. Evaluation of complete identified terms exactly as they appear in the list and using the raw sub-corpus text without tokenization.

Results at this stage were very poor given the nature of the

Arabic text.

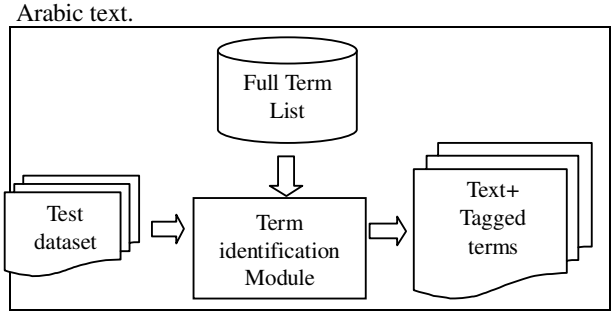

Figure 1. Experiment1-Evaluation Round1

Common challenges in Arabic text tokenization affected drastically the results of the first round of evaluation of term identification for the following reasons (Samy et al., 2006; Habash et al., 2009; Farghaly & Shaalan, 2009):

- **Diacritics.** The unsystematic use of diacritics (vowelization) in Arabic texts is a common feature in Modern Standard Arabic (MSA). In newswire texts it is mostly not used. However in some cases it might be used to avoid some ambiguity. Thus, diacritics can appear for all letters, some, one or none. There are no standards to organize their use.

For example the word "canal" [**قناة**] appeared in the evaluation dataset in different forms:

No-diacritics	قناة	canal
One diacritic "ً" on the first letter	قَناة	canal
Two diacritic "ً" on the first letter and ُ on the last letter	قَناةُ	canal

Table 4. Example of diacritics

- **Enclitics.** The use of enclitics is a basic feature in the morphological nature of the Arabic language. The enclitics are words/letter representing different Parts of Speech but which appear attached to another word forming up one token. For example "del" in Spanish which is the preposition "de" followed by article "el". In Arabic, enclitics can appear as pre-clitics (determinate articles, prepositions, conjunctions) or as post-clitics (possessive pronouns, accusative case endings, etc.). This feature can affect the term identification since terms could appear preceded by an article, a conjunction or a preposition. For example, the following table shows the same term appearing in different forms, each as a different token, due to the use of enclitics.

No-enclitics	التهاب	inflammation
One enclitic "و" conjunction "and"	والتهاب	and inflammation
Two enclitics "وال" conjunction and article "and "the"	والالتهاب	and the inflammation

Table 5. Example of enclitics

In the first round of evaluation, the identification module only identified *24 terms* out of 389, i.e. *6.5%* which is unacceptable result for identification. All the terms identified were correct, i.e., precision is 100%. These poor results led us to carry out a second round of evaluation.

Total of identified terms	24
Total term occurrences	389
Correctly identified	24

Table 4. Results of Experiment 1-Round1

Experiment1-Evaluation-Round 2. In this second round, before running the term identification module using the list of complete terms, the test dataset was normalized and tokenized through a basic tokenization module. In this basic tokenization features such as enclitics and diacritics were normalized, so that the term identification module could run on the normalized tokens.

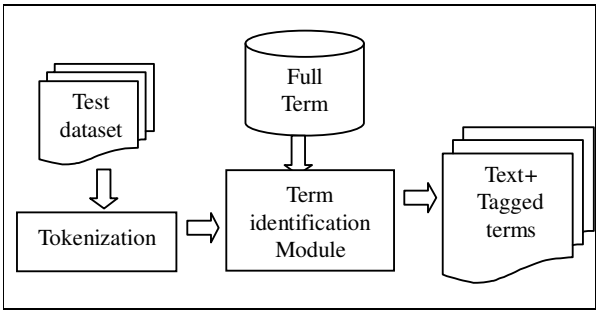

Figure 2. Experiment1-Evaluation Round2

As expected, the results improved significantly after tokenization. 127 more terms were successfully identified after considering enclitics and diacritics raising the overall coverage from 6.5% to 38.9%. This low coverage is justified by the limitation of the list since it only includes some of the medical terms and it is not a comprehensive medical termbase. For example, the list doesn't include pharmaceutical terms nor drug names or chemical components. However, to increase the coverage of term identification, we tried a third round of evaluation.

Total of identified terms	151
Total term occurrences	389
Correctly identified	151

Table 5. Results of Experiment 1-Round2

Experiment 1-Evaluation-Round 3. In this round and to increase the term identification in the test dataset, we changed the search strategy, instead of searching first by the longest multiword terms, we indexed the list of multiword-composite terms into a single term index.

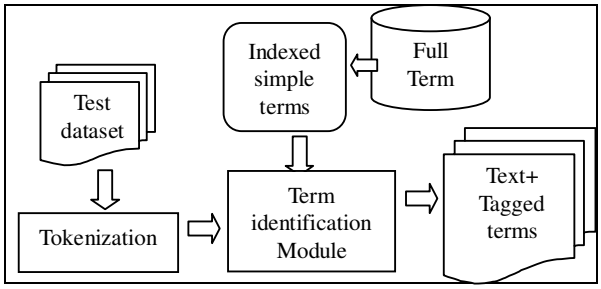

Figure 3. Experiment1-Evaluation Round3

The results of evaluation in round 3 showed an increase in the percentage of correctly identified terms. Out of 666 occurrences of simple indexed terms, 361 were successfully identified raising the overall coverage in the test dataset to 54%.

Total of identified terms	361
Total term occurrences	666
Correctly identified	361

Table 6. Results of Experiment 1-Round3

Error analysis showed that the reasons behind the un-identification of terms, in its majority, were due to the following phenomena:

- The list of terms is incomplete, it only includes 3473 term, so it is not covering all medical domain.
- Pharmaceutical and chemical terms are not included in the term list.
- The high specialization of the term list compared to the text type is one of the challenges in evaluation. The texts of the corpora are not all specialized, most belong to the newswire which is not highly specialized. This is why some words are border line between general domain and specialized domain such as "surgery" or "operation". In this case these words were not included in the term list, although in this context they are counted as terms.
- Some terms were not identified due to features related to the inflectional nature of Arabic. These features include the dual of some words. For example the word "canal" [قناة]is present in the term list, however in the test dataset, it appears several times in dual form as 2 canals[قناتين] and, thus, it was not identified. Another example is 2 eyes [عينتين] or 2 legs[ساقين].
- Some terms were not identified due to morphosyntactic features. In some cases, two characters (one letter and a diacritic) are added to the indeterminate noun in accusative case. For example, the word "disease" [مرضٍ *marad*] in accusative case is [*marad***an** مرضاً]. In other flexional cases in Arabic, the last letter in some adjectives or nouns if "weak letter" [ي، و] might be omitted according to its case. For example, the adjective "infectious" [معدي *mo'dy*] if preceded by a preposition, the last weak letter is omitted and it occurs as [معد *mo'd*]. In these cases, the term is not correctly identified.
- Some orthographic mistakes in the text for example [fats "دهون"*dohoon* appears misspelled as "ذهون"*zohoon*]
- Different dialectal forms or transcriptions of latin name such as "progesterone" which appears in two forms "بروغسترون" and "بروجسترون". Also, "enzyme" might appear as [إنظيم] or [إنريم] with different transcription for /z/.
- Also the compositional nature of Medical terms in Arabic use different syntactic (phrase) structures for example it could use the apposition by *idafa* (2 consecutive nouns the first indeterminate and the second is determinate) or it could use a noun phrase composed of a nucleus noun and a prepositional phrase. For example, in Arabic, the equivalent to low-temperature could appear "انخفاض الحرارة" [low temperature] or "انخفاض في الحرارة" [low in temperature].

Experiment 2-Evaluation-Round 1. In this experiment, we only used the reduced list of 410 Arabic equivalents to Latin suffixes and prefixes. The dataset of Arabic text was not tokenized.

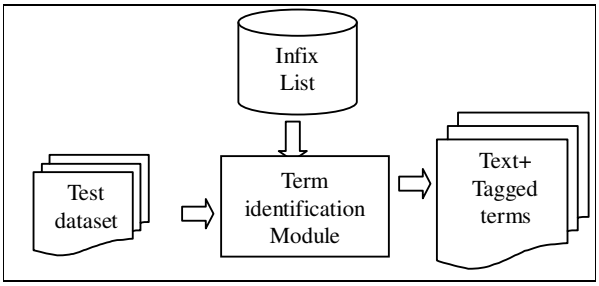

Figure 4. Experiment2-Evaluation-Infix List

Results in this round outperformed its equivalent round in experiment 1, but with lower precision 65.09%.

Total of identified terms	106
Total term occurrences	389
Correctly identified	69

Table 6. Results of Experiment 1-Round3

The advantage of this approach is that is it is not time-consuming neither does it require the effort of maintaining ad validating a list of thousand of terms. However, the disadvantage is the low precision.

Experiment 2-Evaluation-Round 2. To enhance the coverage, the same experiment was performed but after a tokenization phase for the text of the test dataset.

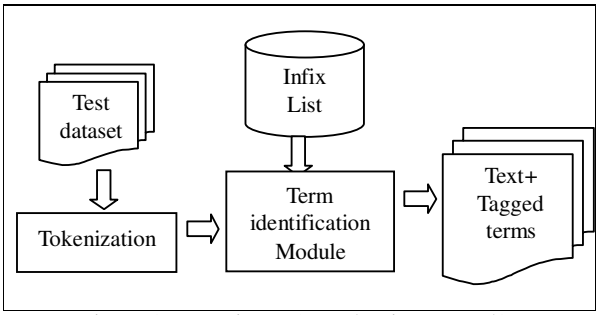

Figure 2. Experiment1-Evaluation Round2

Results at this round of evaluation slightly over performed its equivalent experiment using the full term list achieving a recall of 39.4%, and a precision of 65%.

Total of identified terms	154
Total term occurrences	389
Correctly identified	69

Table 7. Results of Experiment 2-Round2

6. Conclusions and Future Work

Given the above results and evaluation, it is clear that the use of a simple test of 410 terms performed in a satisfactory way compared to a list of 3473. We are aware that both lists are incomplete, however, the effort and time required to extend a full term list is not comparable with the time and effort required to extend a reduced list of infixes. It is also clear that precision is not also the same. However, we could still have a good coverage by a list of infixes.

On the other hand, a corpus of Arabic Medical Text is quite an innovative resource. Nevertheless, the state –of-art of the Arabic language in this domain represent some challenges since texts might not be at the same level of specialization which could convert the process of evaluating how representative the corpus is, into a quite challenging one.

Finally, since the scope of MULTIMEDICA project is multilingual, methodology and annotation are to be applied in comparable corpora in English, Spanish and Japanese aiming at an integrated multilingual platform for terminological and translation purposes as well as for other general purposes.

For future work, the lists of terms are to be extended and features of syntactic structures ruling the composition of Arabic medical terms are to be considered for future work.

7. Acknowledgements

This research has been funded by a grant from the Spanish Government (R&D National Plan Program TIN2010-20644-C03-03)

Special thanks to Ms. Sarah Ahmed Abbas, Spanish Language Department, Cairo University for her efforts in validating the Arabic Term List.

8. References

Chung, W. and Chen, H. (2009). Browsing the underdeveloped Web: An experiment on the Arabic Medical Web Directory. Journal of the American Society for Information Science and Technology, 60 (3), pp. 595-607.

Habash, N.; Rambow, O. and Roth, R. (2009). Mada+tokan: A toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization.. In Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt.

Farghaly, A. and Shaalan, K.(2009). Arabic Natural Language Processing: Challenges and Solutions. ACM Transactions on Asian Language Information Processing (TALIP), Association for Computing Machinery (ACM). TALIP, Vol 8, Issue 4, December 2009, pp.1-22.

Ferguson, Ch. A. (1959). Diglossia. Word 15, pp. 325-340.

Samy, D.; Moreno-Sandoval, A.; Guirao, J.M. and Alfonseca, E. (2006). Building a Multilingual Parallel Corpus (Arabic-Spanish/English). In Proceedings of Language Resources and Evaluation (LREC), Genoa, Italy.