Information Extraction from Arabic Medications Leaflets

Adnan Yahya, Hala Salameh, Maram Belbeisi and Noor Shamasneh
Computer Engineering Department
Birzeit University, Birzeit, Palestine
yahya.birzeit.edu

***Abstract*—Making information in electronic documents easily accessible has been a major concern over the past years. There has been increasing interest in gleaning information from unstructured text and presenting it as structured data using information extraction (IE). Since Arabic has seen major growth in web content, mainly unstructured text, the need for IE from Arabic documents has gained importance. The processing capacity needed for IE far exceeds human ability to extract knowledge manually. The medical field is one such area, where awareness of health issues makes the task of automating medical informatics crucial for better access to medical knowledge. Thus, work on extracting information from medical documents has increased rapidly. In this paper we address the issue of IE from Arabic drug leaflets. We use a combination of rule-based, machine learning and deep learning methods and employ a suit of tools that account for the particularities of Arabic to extract information from Arabic drug package inserts to make this information available in structured form and thus better accessible to regular users and health care providers. A prototype system that utilizes the IE results was developed with useful functionality such as alerting to possible Adverse Drug Reactions (ADR) and finding drug alternatives.**

***Index Terms*—Arabic Information Extraction, Arabic NLP, Health Informatics, Processing Drug Inserts.**

I. Introduction

Text understanding is a fundamental unsolved dilemma in artificial intelligence. Information extraction (IE) is about getting the information in a document and presenting it as database entries or relationships between entities, an important task for grasping the meaning of texts in various fields.

Significant deficiencies have been observed in the task of IE from Arabic texts. Many of the available tools don't adequately support Arabic. Consequently, to put Arabic with the languages machines can process efficiently requires major effort on part of researchers [1].Towards addressing this problem, an IE system, restricted to the domain of Arabic drug leaflets, is developed to turn the brochure free text into easily searchable structured data[1]. This data can then be used for building medical systems that are expandable and easy to use.

The developed system consists of multiple modules. First, the sectioning and labeling module where leaflets are divided into sections, and labels are assigned to each section. Second, the Named Entity Recognition (NER) module to extract entities from each section. In this stage, we use a combination of rule-based, classical machine learning and deep leaning approaches. For each section, whenever possible, we tried to implement all three approaches and adopted the one that gave the best results. Third, the Relation Extraction (RE) module, where entities are connected together to form relations.

The rest of the paper is organized as follows. In section II we give some background material and related work. In section III we outline the general structure of our IE system. In section IV we describe the approaches used and the results achieved for NER. In section V we discuss relation extraction on the example of drug composition, dosage and side effects.In section VI we give a brief description of a prototype system to utilize the extracted data into a useful system for healthcare professionals and ordinary user. Finally, in section VII we give our conclusions and point to possible directions of future work.

[1]The terms "drug" and "medication" as well as the terms "leaflet" and "insert" are used interchangeably in this paper.

II. Background and Related Work

I this section we provide basic definitions, some background material and discuss some of the tools used in IE.

A. Information Extraction and Related Tasks

1) Information Extraction (IE): IE has been extensively studied in various research communities and has its genesis in the NLP communities where it evolves around two fundamental tasks: named entity recognition (NER) and relation extraction (RE). The former refers to identifying entities mentioned in the text and classifying them into groups such as people, organizations and locations, while the latter points out the semantic relationships between the recognized entities. An example would be sideEffectOf (Hypoglycemia, Amarrex), where both Hypoglycemia and Amarrex are extracted NEs and sideEffectOf is the identified relation.

Extraction systems utilize NLP tasks to get information that can help in the recognition of text structure, including tokenizing, part of speech (POS) tagging, among others.

2) Named Entity Recognition (NER): The NER task is isolating and classifying text entities into a predefined set of categories. Solutions to NER can come from manually crafted patterns to form rules to locate and identify entities. Another approach uses machine learning which treats the text as a sequence labeling problem to model many NLP tasks such as POS tagging and Word Sense Disambiguation [2].

3) Relation Extraction (RE): RE detects and characterizes the semantic relationships between the entities defined at the NER stage. For example, the relation that could be extracted from the sentence: *"The drug contains lactose"* is the relation: `Contains(drug, lactose)`.

Relation Extraction can be seen as a classification problem, which classifies the relation between a pair of entities co-occurring in the same sentence into one of the predefined relation types. There are many classification approaches, the most common are feature based and kernel-based classifications [3].

4) Arabic Language and IE: Arabic NLP is a challenging task due to several reasons. First of all, having an undiacritized text adds ambiguity that makes most NLP tasks much harder. Also, Arabic rich inflectional morphology complicates tasks like tokenization, stemming and POS tagging. Moreover, the absence of capitalization makes tasks like NER harder.

B. Related Work

The study of Jian Jing [3], talks about IE and its importance in text mining. Arabic IE has received some attention, even though extracting information form Arabic text still faces some challenges. RelANE [4] discusses the first tool that detects the semantic relations between Arabic named entities. [5] discusses IE from Arabic law documents, characterized by their highly formal language.

IE in the medical domain aims to convert medical reports into structured information so that the information can then be analyzed, aggregated, and mined. IE from Medical Notes [6] is a study that discusses several methods that automatically identify drug, dosage, and method of delivery from transcribed physician notes. Extraction of Adverse Drug Effects from Clinical Records aimed at extracting adverse drug reactions from the electronic health records (EHR) [7]. Multiple studies went into using NLP to identify Pharmacokinetic Drug-Drug Interactions described in Drug Package Inserts [8] and discussed using machine learning algorithms to identify Pharmacokinetics in package inserts(PI) using a corpus consisting of annotated statements collected from FDA [9].IE from drug labeling for product-specific guidance assessment is given in [10]. Another use of IE in the medical domain is to extract information from medication leaflets such as PharmInx which extracted dosage from Portuguese medication inserts [11], something comparable to some of what we try to do in this paper. Much more drug data is available in English than in Arabic. This is the case for ADR reports [9], drug databases [12], [13] and more. That's why we utilize the data in a cross lingual manner to improve the Arabic extraction process.

C. Tools

Table I summarizes the tools we used to implement the different tasks of our IE system.

TABLE I
TASKS AND TOOLS USED

Task	Tools Used
Tokenization	Standford [14]
Stemming	NLTK [15]
POS Tagging	Madamira [16], Standford [14]
Medical Language System	UMLS API [17], PubChem [18]
Classical Machine Learning	Scikit-learn [19]
Spell Checking	Google Spell Checker [20]
Translation	Google Translate [21]
Deep Learning	DeepCRF [22]

III. Information Extraction System Description

A. Overview

Our system focuses on IE from Arabic drug leaflets. It has three modules: the first consists of preprocessing steps such as normalization, sectioning and section labeling. The second identifies the entities in the text, classifies them into predefined categories with the help of POS tagging and other useful NLP results. In the third module relations between the recognized entities are then identified mainly through the section labels. Figure 1 explains the stages we followed.

The work was done using a dataset that we collected from local manufacturers and drug importers.

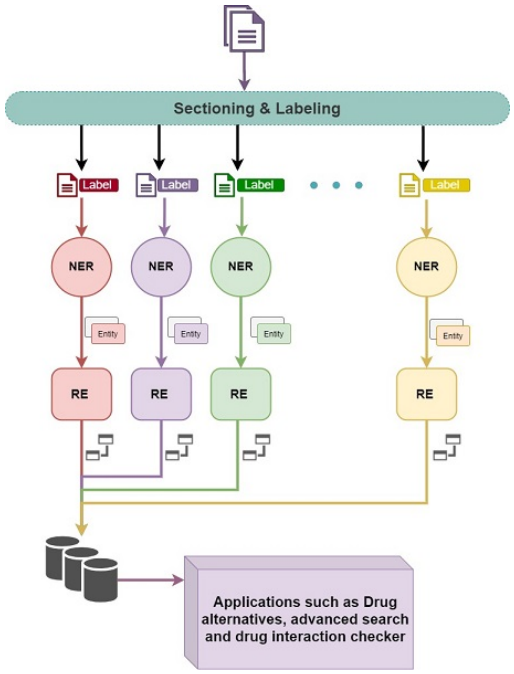

Fig. 1. Basic Modules for IE from Drug Leaflets

B. The Dataset

Our dataset is a collection of Arabic leaflets for drugs from the four major Palestinian pharmaceutical companies and imported drugs. We worked with 397 leaflets from 4 local manufacturers (50:BPC, 51:Jerusalem, 110:Beit Jala, 90: Dar Al-Shifa) and drug importers (96).

The dataset elements needed to be put in a form suitable for further text processing and that was done in two stages: Stage I to Convert PDF to MS Word format, sometimes manually, then to XML format. Using XMLlib [23] to enable reading inserts tables and text boxes was done in Stage II.

C. Preprocessing

The goal of text preprocessing is to prepare the leaflet for the extraction process. It consisted of the following steps:

1) Normalization: to unify frequently misspelled, so called confusion letters ه ـ ة ـ ي ـ ى ـ آأإا.
2) Removing diacritics.
3) Removing stop words.

D. Sectioning and Labeling

We experimented with IE from the full documents but the results were not encouraging. We found that dealing with individual sections always yields much better results.

We analyzed a large number of leaflets and identified the keywords that are needed to determine the individual sections of the leaflet (using a rule-based approach). The leaflets were divided into blocks, bounded by new lines and classified based on the presence of keywords specific to a particular section, then each block was assigned one of the 21 section labels.

E. Evaluation Metrics

To evaluate the different parts of the system, the precision (P), recall (R) and f-measure metrics were used. To compute these measures, in the rule-based approaches we used cosine similarity between the manually annotated entities and the extracted ones using an experimental threshold. Similarity exceeding the threshold results in positive classification. For the ML tasks, we used 70% / 30% data split.

IV. Named Entity Recognition (NER)

The input for this stage is the labelled sections and the output is the NEs in these sections. Even though 21 sections were obtained in the sectioning stage, we worked on 6 sections only due to types of relations we were interested in. Three approaches were used, Rule-based, classical Machine Learning and Deep Learning. For each section, one or more approaches were tried and the one with the best results was selected.

A. Rule-based NER

Here we use manually crafted rules to extract entities with the help of NLP tools output such as POS tags. Next we discuss rule-based NER from different sections of the leaflets. The rule crafting was based on a limited number of leaflets while the testing was applied to the entire set of leaflets. Next we give the details.

1) Composition: To extract the chemical substances entities from the leaflets, only the composition and introduction sections are considered. We implemented the rule-based approach in two stages: POS tagging using MADAMIRA [16], followed by rules to extract the chemical substances and their quantities. The rules were based on the closest distance between units, chemical substances and their associated quantities. The recall was 54.8% while the precision was 71.3% and the f-measure was 62.0% for composition extraction.

2) Therapeutic Group: Therapeutic group is the class of medicines that the drug belongs to. The Therapeutic section was the input and the output was the set of Therapeutic groups entities to which the drug belongs. A two stage rule-based method was used. The preprocessing stage included removal of stop words and common phrases. Additionally, two lists - one in English and one in Arabic- containing all the therapeutic groups, were collected from the international drugs.com site [13]. The second stage is crafting a set of rules to extract the `Therapeutic group` entities with the help of the two lists and Google Translate service. The algorithm works as follows:

1) Each Arabic phrase representing an extracted entity is translated to English using Google Translate and spell checked.
2) The resulting English phrase is compared with the English therapeutic groups list using edit distance. If there is a match from the list, the Arabic corresponding class is added to the result. Otherwise we go to next rule.
3) Using Google, we search for the English sentence in drugs.com site [13], and get the first result that contains "drug-class" phrase in its link, extract the drug class name from the link and match it with the list. If no match is found go to the next step.
4) Search in international drug bank site [12] for drug composition and get the classes of the drug compositions. Search for those classes in drugs.com site in order to unify them with our list. This step always returns results since drug composition is used, not the text itself.

The extracted Therapeutic groups entities were connected to the drug with `BelongTo(Drug, Therapeutic group)` relation. The evaluation results for this process were 79%, 86% and 82% for precision, recall and f-measure, respectively.

3) Side Effects: The Side Effects section is the input for this extraction process. After preprocessing, several lists were compiled: the first has the used medical terms,the second has the human body parts and the third contains the side effects words/phrases that do not need explicit naming of human body parts such as constipation إمساك, diarrhea إسهال . The procedure then proceeded as follows:

1) In the Side Effects section, search for the first word that indicates the start of the side effects list. Select the text between those words to the end. If no such word is present, the entire section is selected.
2) Split the selected text into sentences using punctuation marks ". , ; :"
3) Stem words in each sentence and search for medical terms and human body part words/phrases in the stemmed sentence and discard any sentence that doesn't include words from these lists.
4) Remove some common non side-effect words that do not add useful information from the filtered sentences such as دواء (drug), treated as domain specific stop words.
5) Tokens are POS-tagged using Stanford POSTagger [14].

6) Each sentence is checked for the occurrence of the prepositions-(and) و- and -(or) أو -. If none is found, the sentence is considered a side effect. If found, the sentence may be indicating multiple side effects and is split into two or more sentences, based on the linguistic properties of the tokens in the sentence.

The evaluation results for the side effects extraction were 78.2% precision, 89.4% recall and 83.1% f-measure.

4) Drug Interactions: Leaflets state the drug interactions in different forms; some mention specific drug names, while others may give a therapeutic group of the drug. To extract drug interactions using a rule-based approach, first the stop words and unneeded words such as:

نجاعة -efficacy- and إبلاغ - reporting- were removed . Next, we applied the following rules:

1) Determine the start of the interactions list by searching for specific keywords defined manually such as التالية (the following), المتزامن مع (simultaneously) الآتي , ...etc in the drug interaction section.
2) The list is divided into possible interactions using the punctuation marks and Arabic connectors as delimiters.
3) Each potential interaction is checked for being a real interaction by applying the following tests, in that order:
   a) Test 1: UMLs [17] with Google Translate then Search the UMLs for the English translation then get the UMLs category for the first search result. If it was any of: Biologically Active Substance, Pharmacologic Substance, Element, Ion, Isotope or Organic Chemical, an interaction is confirmed.
   b) Test 2: UMLs with search in WebTeb site [24] translation and return the first three English terms in the first article returned. Repeat steps of Test 1 for the three retrieved English terms.
   c) Test 3: Arabic Wikipedia Perform a google Search with the possible interactions text plus the word ويكيبيديا (Wikipedia) as the query. If no link is returned, the test fails. For the first Wikipedia page search for "PubChem", "ChEBI" or "ATC" in the text. If any is found,the test succeeds else it fails.
   d) Test 4: English Wikipedia: Translate the text to English using Google Translate and repeat test 3 for the English translation.

If the possible interaction passes any of the four tests, it is added to the final interactions list. Otherwise, it is rejected.

The evaluation results for drug interactions extraction were 79.1% precision, 88.3% recall and 83.2% f-measure.

5) Manufacturer: We extracted the manufacturer name using only the Company section using Rule based approach. The evaluation results for drug manufacturer were 93.1% 91.4% and 92.3% precision, recall and f-measure, respectively.

B. Machine Learning NER

The machine learning (ML) approach was used when the rule-based approach didn't give acceptable results for a given section. The common word features in all sections were the word text, POS tag, is-a-digit (yes or no) and word location in the sentence. Individual sections may have own sets of additional tags. An IOB (Inside–Outside-Beginning) tagging was used to handle phrases consisting of more than one word. I-prefix indicates that the token is inside a chunk, B-prefix indicates that the token is the beginning of a chunk and O-prefix indicates that the token belongs to no chunk (outside).

We experimented with six classifiers: Decision Tree, SGD, SVM, Random Forest (RF), Perceptron and CRF. The selection was based on work done by other researchers and our own preliminary experiments.

1) Composition: The dataset for the Composition model consisted of the Introduction and Composition sections of each leaflet, with 1048 sentences and 12800 words overall. The NEs that the model recognized were the drug name, chemical substances, quantities and units. Table II summarizes the results for different classifiers and shows that the highest F-measure was 87.4% using manual POS tagging and CRF classifier. When we applied CRF with automatic POS tagging using Madamira [16], the corresponding value decreased to 82.5%, possibly due to the medical terms in texts, where Madimara [16] is not specialized in the medical context.

TABLE II
MACHINE LEARNING PERFORMANCE FOR COMPOSITION: WITH MADAMIRA AND MANUAL POS TAGGING

Classifier	Precision	Recall	F-measure
Decision Tree	77%	72%	74.4%
SVM	79%	73%	75.9%
SGD	69%	68%	68.5%
Random Forest	79%	62%	69.5%
Perceptron	67%	69%	68.0%
CRF	91%	84%	87.4%
CRF+Madamira POS Tagging	82%	83%	82.5%

2) Drug Interactions: The dataset used had 458 sentences and 12367 words from Drug Interactions sections of the leaflets. We followed the same steps as in Composition ML NER. The only entity the model recognized was the Drug Interactions: (B-INTERACTION and I-INTERACTION), where each leaflet usually contains 10 or more interactions. Table IV summarizes the results. Again, CRF gave the highest F-measure (78.7%).

C. Deep Learning NER

Deep learning (DL) was used in extraction when rule-based and machine learning approaches performed poorly. The model used was Deep CRF [22]: an end to end sequence labeling that uses neural networks architecture with a combination

TABLE III
MACHINE LEARNING PERFORMANCE FOR DRUG INTERACTIONS

Classifier	Precision	Recall	F-measure
Decision Tree	68%	61%	64.3%
SVM	75%	69%	71.9%
SGD	75%	55%	63.5%
Random Forest	74%	51%	60.4%
Perceptron	69%	63%	65.9%
CRF	83%	75%	78.7%

of bidirectional LSTM, CNN and CRF. CRF has proven ability to deal with sequential text and extract named entities [22].

1) Composition: The same dataset used in the classical ML was used for DL, with some additional tags to identify more entities. The additional annotation was done after noting that the results from DL were higher than from classical ML and rule- based approaches. The newly recognized entities included non-active substances used for preservation or coloration or to add taste. Table IV gives the tags used with explanations and examples.

TABLE IV
COMPOSITION NER TAGS FOR DEEP LEARNING

Tag	Explanation	Tag	Explanation
CHEM	The name of the chemical substance	DCHEM	Description of the chemical substance
NUM	The number part of the chemical substance quantity	UNIT	The unit part of the chemical substance quantity
FOR	Grouping the compositions	EQU	composition: chemical substance, quantity= another substance
NECHEM	Number part of the inactive chemical substance quantity	NENUM	Number of times the dosage per period
NEUNIT	Unit part of inactive substance quantity		

Using Deep CRF, three models were built for: active, inactive and both substances. The last model scored best for many of the tags as in Table V, Table VI and Table VII.

TABLE V
ACTIVE SUBSTANCES EXTRACTION PERFORMANCE

Tag	Precision	Recall	F-Measure
DRUG	90.6 %	76.8 %	83.1 %
CHEM	83.3 %	88.5 %	85.8 %
UNIT	92.1 %	97.3 %	94.6 %
NUM	90.2 %	99.6 %	94.7 %
DCHEM	67.3 %	64.7 %	66.0 %
EQU	81.3 %	100 %	89.7 %
FOR	100 %	20 %	33.3 %
Weighted Overall	87.6 %	89.6 %	88.6 %

2) Drug Interactions : We used the same dataset as for classical ML. The DL model gave 86% precision and 88.8% recall with 87.4% F-measure, higher than the rule-based and the best results for traditional ML.

3) Dosage: Dosage refers to the amount of drug that should be taken with a specific frequency over a specific interval.

TABLE VI
INACTIVE SUBSTANCES EXTRACTION PERFORMANCE

Tag	Precision	Recall	F-Measure
DRUG	93.1 %	71.1 %	80.6 %
NEID	85.7 %	46.2 %	60.0 %
NECHEM	47.1 %	53.3 %	50.0 %
NENUM	100 %	42.9 %	60.0 %
NEUNIT	75.0 %	42.9 %	54.6 %
Weighted Overall	82.9 %	65.1 %	72.9 %

TABLE VII
COMBINED ACTIVE AND INACTIVE SUBSTANCES EXTRACTION PERFORMANCE

Tag	Precision	Recall	F-Measure
DRUG	92.7 %	75.6 %	83.3 %
CHEM	82.1 %	91.6 %	86.6 %
UNIT	93.6 %	97.3 %	95.4 %
NUM	90.9 %	98.2 %	94.4 %
DCHEM	73.3 %	64.7 %	68.7 %
NEID	64.0 %	53.3 %	58.2 %
NECHEM	64.0 %	53.3 %	58.2 %
NENUM	100 %	28.6 %	44.5 %
NEUNIT	80.0 %	57.1 %	66.6 %
EQU	81.3 %	100 %	89.7 %
FOR	100 %	10 %	18.2 %
Weighted Overall	87.6 %	87.5 %	87.5 %

To extract the dosage from the leaflet, different entities have been extracted using DL models. The data-set consisted of the dosages sections from 400 leaflets, with 30500 words. 15 tags were extracted using the IBO system. Among these tags are: *SRANGE and ERANGE used to describe the MIN and MAX dosage when the medicine is in liquid condition, and **DSRANGE and DERANGE used to describe the MIN and MAX dosage when the medicine is solid. The DL model gave 76.8% precision and 67.7% recall with 71.9% F-measure.

D. Summary of NER Results

We evaluated the NER task with 30% of the dataset for testing. For some sections, we implemented the rule-based, classical machine learning and deep learning approaches and the approach that gave the best result was chosen. However, in some cases, such as the Side Effects, only the rule-based approach was implemented due to the fact that side effects were really long and had large variations. For the Dosage section, only deep learning was used since the corresponding section was too complex for rules. The best results achieved and the approach used are summarised in Table VIII.

TABLE VIII
BEST RESULTS FOR ENTITY EXTRACTION FROM MEDICAL LEAFLETS

Entity Extracted	P	R	F	Tested	Best
Sectioning/Labeling	99.0%	94.4%	96.9%	R	R
Composition	87.6%	87.5%	87.5%	R,ML	ML
Side Effects	78.0%	89.0%	83.1%	R	R
Drug Interactions	86.0%	88.8%	87.4%	R,ML,DL	DL
Therapeutic Group	79.0%	86.0%	82.4%	R	R
Manufacturer	93.6%	91.2%	92.4%	R	R
Dosage	76.8%	67.7%	71.9%	DL	DL

It is interesting to note that classical ML didn't make it into the "Best" column for any of the entities. It seems that Rule-

based approach worked well for simpler classifications such as Sectioning and Manufacturer and DL worked well for more complex cases such as Composition and Drug Interactions.

V. Relation Extraction

After NER we had to join the extracted entities into relations, as the final stage of our IE from drug leaflets. The relations identified are strongly related to the sections from which they were extracted: some sections have more than one relation and others, such as drug interactions, have only one. Moreover, the sections where the rule-based approach was used, in most cases had the relation extraction embedded within the NER and the relation consisted of the extracted NE and drug name joined into a relation defined by the section name. Some sections needed additional computations beyond NER for relation identification such as the side effects where synonyms and complex relations existed which will be discussed in details below. Extra relations could be identified using the extracted entities and base relations, such as Drug-Drug interactions that can be identified by seeing which substances and therapeutic groups the drug interacted with and from there identify the Drug-Drug interactions. Next we elaborate on the cases where substantial work was needed to extract relations (or relation instances).

A. Composition

After extracting the NEs from the Composition section using the DL model, a code was written to put these NEs into a structured relation. In order to group the tags into sentences, a set of rules were applied to the NER model output, as follows:

1) All the B-TAGS and the accompanying I-TAGS are concatenated together to form entities.
2) CHEM entities -Chemical substances- are concatenated with the DCHEM entities -Chemical substances description- closest to them.
3) A composition is created using the CHEM entity with the closest UNIT and NUM entities, if present.
4) If an EQU entity is found, the composition existing after the EQU tag is taken instead of the one before it since we observed that the quantity always comes after the EQU tag that is related to the composition.
5) If a FOR entity is encountered, the compositions until the next FOR entity are added to that FOR entity in order to group the compositions.
6) The same procedure is followed with inactive substances but with no EQU or FOR tags since with inactive substances, the compositions are neither grouped, nor found in equality forms.
7) The extracted entities are added to their respective database tables. In addition, the composition entity is translated using Google Translate and google Spelling. After that, the UMLS code returned by the UMLS API and the English translation are added to the database. This step was done to make the tables in the database bilingual and for future use in services provided in the prototype system such as finding drug alternatives.

B. SideEffects

The extracted side effects are semantically normalized so that, for example, صداع,وجع في الراس , الم في الراس (various ways of expressing "headache" in Arabic) would all be identified as synonyms and normalized to the same phrase. The strategy for that is summarized in the following steps:

1) Translate the extracted side effect from Arabic to English using Google Translate.
2) Using UMLS API [17], try to retrieve the code for the English translation of the extracted side effect. This code represents the unified side effect in UMLS database. For example, both "Headache" and "Head pain" have "C0018681" code in UMLS.
3) If the code already exists in the database, then only the Arabic extracted side effect is added as a synonym and is linked to the existing UMLS code and to the drug. The English translation is linked to as a synonym.
4) If the UMLS code is not found in the database, the category for the received code is fetched, to check if the retrieved category belongs to the predefined set of side effect categories ('Disease or Syndrome', 'Sign or Symptom') and the code is added to the database as a new side effect with the English translation as synonym, as in the previous step .
5) If no code is retrieved from the UMLS, a request is sent to Google containing a query of "side effect + ويكيبيديا (Wikipedia)" and the first obtained result from Wikipedia that has a category of بوابة طب(Medical Gate) [2] is returned. If the result from Wikipedia exists in the database only a new synonym is added, otherwise, both a new side effect and synonym are added.
6) If no results are retrieved from UMLS or Wikipedia, the side effect is added as is, with no synonym or code.

C. Dosage

Steps were applied to the DL model results to obtain the posologies for each drug. Each posology consists of relations between elements like: dosage unit, value, frequency…etc. These steps can be summarized as follows:

1) Combine "B-TAG"s with the corresponding "I-TAG"s for Dosage entities.
2) The range elements are split to get max/min components.
3) SRANGE with the subsequent ERANGE are linked to construct a range element with min and max.
4) DSRANGE and the subsequent DERANGE are used to get dosage and range elements based as follows:
   a) Translate DSRANGE and DERANGE to English.
   b) Search for number words such as (two, eleven…) in the DSRANGE translation and match those words with numbers (2, 11...etc). If no number words are extracted, value defaults to 1.

[2] بوابة طب the main category that refers and contains medical subjects in Wikipedia, used to categorize medical articles at the end of the page with other sub-categories to be identified as medical content.

c) Repeat the previous step for DERANGE.
d) Construct a range element from the numbers extracted, where the number extracted from DSRANGE and DERANGE represent the min and max components, respectively.

5) Each VALUE (Dosage quantity) element is connected to nearest DOSAGE element.
6) Each MATH "Relation" element (Dosage as a function, say to body mass) is connected to seen previous dosage
7) FREQ "PER + FREQ"[3], STOP and DUR are connected to nearest DOSAGE into one DOSAGE entity.
8) Each CASE and GRP (group, say children) is connected to the nearest Dosage to obtain posology.

The complex relation extraction was not formally evaluated due to the absence of any annotated dataset or API for that.

VI. A Prototype Application

Using the extraction results we developed Dawa'y Tech (MyMedicine Tech) application that focused on increasing health awareness and medical information accessibility, targeting both patients and health care workers. It mainly provides drug related information in Arabic, but is also available in English. One possible use scenario for the application is to scan the leaflet of a drug and extract the entities/relations therein to find more relevant information about that drug. This may be viewed as another stage of the IE process whereby we define new relations between drugs or drug components not readily available in the leaflet. An Arabic speaking user may use that to alert his doctor on the possible problems of a prescribed medicine and the doctor may use it to prescribe alternative drugs for problematic cases. For space limitations we do not elaborate here on the functionalities currently offered by our application that include:

1) Adverse Drug Reaction (ADR)
2) Drug Alternatives
3) Drug Interaction Checker
4) Drug Name Transliteration

VII. Conclusions and Future Work

We aimed to extract information from Arabic medication leaflets to provide structured data that can be used in useful applications to enable users to access drug data easily. We proved that Arabic information extraction is possible in the sensitive medical field and can be trusted in other fields. The results proved to be of practical value and useful for the community, due to the lack of Arabic extraction systems, especially in the medical domain.

We faced many limitations such as the limited data-set and reluctance of drug companies to provide the needed data in machine readable form. Moreover, due to the scarcity of ANLP tools in the medical domain, we had to turn to English medical NLP tools such as UMLS. To use these tools, translation into English was a must and translation mistakes sometimes caused the extraction to go askew. Furthermore, the performance of common NLP tools frequently failed in the medical domain. To help close that gap we are making the collected dataset and the annotated data available for other researchers. Cumulative errors were faced due to the sequential nature of our implementation: the output of the sectioning goes into NER which finally goes into RE and any obstacles along the way affect the following stages and the final output.
As for future work, we would like to check the applicability of the developed tools to other domains.

[3]PER + FREQ term used to define the prescribed dose during the period of taking the medicine, e.g. 3 capsules per day, means one pill every 8 hours

References

[1] A. Al-Zoghby, A. Ahmed, and T. Hamza, "Arabic semantic web applications – a survey," *Journal of Emerging Technologies in Web Intelligence*, vol. 5, pp. 52–69, 2013.
[2] S. Sunita, "Information extraction," *Foundations and Trends® in Databases*, vol. 1, pp. 261–377, 2008.
[3] J. Jing, "Information extraction from text," *Mining text data*, pp. 11–41, 2012.
[4] I. Boujelben, S. Jamoussi, and A. Ben, "Relane: discovering relations between arabic named entities," *International Conference on Text, Speech, and Dialogue. Springer, Cham*, pp. 233–239, 2014.
[5] W. S. Samah Abu Shamma, Aseel Ayasa and A. Yahya, "Information extraction from arabic law documents," *The 14th IEEE Internationl Conference Application of Information and Communication Technologies (AICT2020). 07-09 Oct 2020 — Tashkent, Uzbekistan*, 2020.
[6] S. Kraus, C. Blake, and S. West, "Information extraction from medical notes," *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems. IOS Press*, p. 1913, 2007.
[7] E. Aramaki, Y. Miura, M. Tonoikeb, T. Ohkuma, H. Masuichi, K. Waki, and K. Ohe, "Extraction of adverse drug effects from clinical records," *MedInfo 160*, pp. 739–743, 2010.
[8] R. Boyce, G. Gardner, and H. Harkema, "Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts," *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 206–213, 2012.
[9] "What we do," 2019. [Online]. Available: https://www.fda.gov/about-fda/what-we-do
[10] Y. Shi, P. Ren, Y. Zhang, X. Gong, M. Hu, and H. Liang, "Extraction from fda drug labeling to enhance product-specific guidance assessment using natural language processing," *Frontiers in Research Metrics and Analytics Vol. 6*, 2021.
[11] B. L. Aguiar *et al.*, "Information extraction from medication leaflets," Ph.D. dissertation, the engineering faculty of the University of Porto, 2012.
[12] "Drugbank," 2019. [Online]. Available: https://www.drugbank.ca/
[13] "drugs.com," 2019. [Online]. Available: https://www.drugs.com/
[14] "The stanford natural language processing group," 2019. [Online]. Available: https://nlp.stanford.edu/
[15] "Nltk," 2019. [Online]. Available: https://www.nltk.org/
[16] A. Pasha, M. Al-Badrashiny, M. Diab, A. E. Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow, and R. Roth, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," *LREC*, vol. 14, pp. 1094–1101, 2014.
[17] "Unified medical language system," 2019. [Online]. Available: https://www.nlm.nih.gov/research/umls/index.html
[18] "Pubchem," 2019. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/
[19] "Scikit," 2019. [Online]. Available: https://scikit-learn.org/stable/
[20] "Google spell checker," 2019. [Online]. Available: https://github.com/adamryman/google-spell-check
[21] "Google translator," 2019. [Online]. Available: https://github.com/adamryman/google-spell-check
[22] "Deepcrf," 2019. [Online]. Available: https://github.com/aonotas/deep-crf
[23] "Xml processing modules," 2019. [Online]. Available: https://docs.python.org/3/library/xml.html
[24] "Webteb," 2019. [Online]. Available: https://www.webteb.com//