CasANER: Arabic Named Entity Recognition Tool

Fatma Ben Mesmia, Kais Haddar, Nathalie Friburger and Denis Maurel

Abstract Actually, the Named Entity Recognition (NER) task is a very innovative research line involving the process of unstructured or semi-structured textual resources to identify the relevant NEs and classify them into predefined categories. Generally, NER task is based on the classification process, which always refers to the previous categorizations. In this context, we propose CasANER, which is a system recognizing and annotating the ANEs. The CasANER elaboration is based on a deep categorization made using a representative Arabic Wikipedia corpus. Moreover, our proposed system is composed of two kinds of transducer cascades, which are the analysis and synthesis transducers. The analysis cascade, which is dedicated to the ANE recognition process, includes the analysis, filtering and generic transduces. However, the synthesis cascade enables to transform the annotation of the recognized ANEs into an annotation respecting the TEI recommendation in order to provide a structured output. The implementation of CasANER is ensured by the linguistic platform Unitex. Then, its evaluation is made using measure values, which show that our proposed system outcomes are satisfactory. Besides, we compare CasANER system with a statistical system recognizing ANEs. The comparison phase proved that the results obtained by our system are as efficient as those of the statistical system in the recognition and annotation of the person's names and organization names.

F.B. Mesmia (✉)
Laboratory MIRACL, Multimedia InfoRmation Systems and Advanced Computing Laboratory, University of Tunis El Manar, Tunis, Tunisia
e-mail: fatmabm@ymail.com

K. Haddar
Laboratory MIRACL, Multimedia InfoRmation Systems and Advanced Computing Laboratory, University of Sfax, Sfax, Tunisia
e-mail: Kais.Haddar@fss.rnu.tn

N. Friburger · D. Maurel
LI, Computer Laboratory, University François Rabelais of Tours, Tours, France
e-mail: nathalie.friburger@univ-tours.fr

D. Maurel
e-mail: denis.maurel@univ-tours.fr

K. Shaalan et al. (eds.), *Intelligent Natural Language Processing: Trends and Applications*, Studies in Computational Intelligence 740,
https://doi.org/10.1007/978-3-319-67056-0_10

Keywords CasANER · Arabic named entity · NER
Transducer cascade · TEI

1 Introduction

Since the MUC (Message Understanding Message conference) conferences, the named entity recognition (NER) has been considered as a sub-task of the Information Extraction (IE) main task. So far, the NER has still been a very innovative research line, which involves processing unstructured or semi-structured textual resources, on one hand, to identify the relevant NEs and on the other hand, to classify them into predefined categories on other hand. The NER task is evolving to realize many objectives, namely enhancing the NLP-application performance. Therefore, recognizing NEs and assigning them to the adequate categories offer an opportunity to create or enrich NE electronic dictionaries and realize the Entity Linking (EL) task. Besides, the NER participates in a large part to enrich documents, in their semantic level that can index Question Answering systems. Furthermore, recognizing and annotating NEs can be a preprocessing to extract SRs between them and increase the search engine efficiency in order to provide relevant responses.

Generally, the NER task is based on the classification process, which always refers to the previous categorizations. However, the NER can encounter many challenges. The first challenge concerns the NE representation that makes its delimitation a complex task. Furthermore, an NER process requires a clear definition to guess the NE limits. The identification phase can suffer from the absence of indicators that may precede an NE. This NE can be similarly enunciated through expressions where the problem of choosing the adequate context arises. Besides, this process may envisage the ANE imbrication, which is very sensitive in its treatment. The second challenge appears after the NE delimitation. This means that the identified NE must be assigned to the adequate category. Nevertheless, the NE categorization is not easy. Moreover, we must determine the best categories to describe the recognized NEs. In addition, the categorization process also contributes to the increase of the granularity level. In fact, an NE can belong to different categories so the ambiguity problem appearing in this case. According to the Arabic NER, the complexity of this language also arises various problems. The main one is its complex morphology due to its agglutinating nature. The complex morphology is highly related to the ambiguity problem. The last challenge is related to the ANE annotation. This annotation must respect a norm that clearly represents the ANE components. Likewise, the ANEs must be described through significant tags having attributes allowing the specification of their categories and sub-categories.

In this context, we exploit the representative Arabic Wikipedia corpora (study and test). We also suggest a deep and detailed ANE categorization to construct a developed hierarchy. In addition, we propose the CasANER system to recognize and annotate the ANEs. Our ANE recognition process is based on (Finite-State) transducers, which couple the recognition and annotation processes. We will exploit generic transducers, which is a new notion improving the recognition transducers. Then, we generate a cascades calling the established transducers in a predefined passage order. The annotation process is based on the TEI recommendation to detail the ANE components.

Our work originality consists in exploiting corpora (study and test) containing articles extracted from Arabic Wikipedia especially form several Arabic countries. In fact, this variety of articles enable us to treat the different styles of the Arabic language, to contemplate the regional writing, such as in the trigger words like “ولاية”,“محافظة” “مدينة” and “قضاء” to introduce a city name and to envisage the different civilizations, such as the use of various calendars (Syriac, Muslim and Gregorian) to describe dates. Moreover, we refine the categorization, which participates in a large part to improve the realization of other NLP-applications. Similarly, our work originality resides to generate using the TEI recommendation a structured output able to be integrated in the semantic Web and to enrich electronic NE dictionaries.

The present paper is composed of five sections. Section 2 presents the state of the art on NER systems based on different categorizations, approaches and domains. Section 3 consists in describing our linguistic study made to identify the ANE categories and forms from our Arabic Wikipedia corpus and describe the different relationships with the ANEs. Section 4 details our proposed method to elaborate CasANER system to recognize ANEs and to annotate them using the TEI recommendation tags. The implementation ensured by the linguistic platform Unitex and the evaluation are presented in Sect. 5. the linguistic platform Unitex is used in this phase. Finally, we give a conclusion and some perspectives.

2 State of the Art on NER Systems

The establishment of the NER systems is a process composed of three fundamental steps. The first step is related to the NE definition, which facilitates its determination. Then, the second step is based on the first one. In other words, the NE definition enables to guess its category and therefore an NE categorization step arises. Furthermore, the third step concerns the choice of the approach and the associated formalism or techniques, which respond to the future system needs. In the current section, we will present the previous NE categorization and some previous reasearch to elaborate NER systems using the different domains and corpora nature.

2.1 Previous NE Categorization

The NER systems revolve around the NE categorization, which is a crucial step for many NLP-applications. It should be noted that the NE categorization is a process aiming to provide adequate NE representation. Furthermore, this process helps elaborate an appropriate hierarchy translating the corpora richness. The NE categorization was proposed for the first time in MUC-6 [20]. In fact, the same MUC-6 categorization was taken up at MUC-7 [13]. However, a new category was added depending on the corpus nature and it appends a slight modification to the last categorization. As well, many conferences and projects adopted the MUC-6 categorization, which is refined in each one based on different corpora related to specific domains and languages. Among these conferences, we quote the two MET (Multilingual Entity Task) conferences that were organized parallel to MUC-6/7. The MET provided a new opportunity to evaluate the NER task progress in Spanish, Japanese and Chinese [25]. Even, the IREX (Information Retrieval and Extraction) is a Japanese project reposing also on the MUC-6 categorization and it allows the adding of the sub-categorization notion [20]. In the CoNLL (Conference on Natural Language Learning), the authors proposed some changes to the MUC categorization by adding a new category called Miscellaneous for the NEs having no determined category. In addition, many evaluation campaigns also adopted the MUC principle, such as the ACE (Automatic Content Extraction) campaigns [15], Ester [18] and annotation models, such as Quaero [19, 5]. The already mentioned categorizations aimed not only at adding new NE categories but also at refining the existent one. At the same time, they set other objectives, such as providing structured and accessible corpora and enhancing the NER task. In the following section, we will present the NER approaches and the associated previous work.

2.2 NER Approaches and Systems

The NER systems are always based on the three main approaches (symbolic, statistical and hybrid) to recognize the NEs and annotate the recognized NEs through a norm, which represents and details their components [8, 33]. Nevertheless, these systems used undefined markup, which responds to their needs and reliability. In what follows, we will present the elaborated NER systems and tools based on the main approaches.

Symbolic approach. The symbolic approach for NER is based on formal local grammars described by hand-crafted rules, which identify the NEs. These hand-crafted rules can be modeled using regular expressions or finite state transducers. In this context, we quote the PERA system developed by [30], which recognizes Person's Names in Arabic scripts. Based on the same PERA functionalities, a NERA system was elaborated by [31]. This system can recognize ANE of 10 categories. In addition, [27] proposed a novel methodology to improve NERA system by identifying its weakness. The improved system is called NERA 2.0, which ameliorates the coverage of the previously misclassified NEs. This enhancement makes the system achieve more reliable results. Using the rule formalism, [2] developed also a system for an Arabic person's name composed of four main rules composing the core system. In [3], the proposed system is dedicated to recognize an event, time and place expressions through a set of general rules: two rules for time expressions and a rule for place expressions. In addition, [12] proposed a tool for both the Part of Speech (PoS) tagging and NER for the Arabic language. According to the NER, the authors elaborate a NE detector, which acts on the text by giving the adequate labels. The elaborated NE detector starts by reading the data set, which are lists containing the person, location and organization names. Then, it splits the input text into words to apply the established rules. Consequently, if the divided words match these rules, then the labels will be assigned, and else the label will be unknown. Moreover, the authors used three labels, which are Person (PERS), Location (LOC) and Organization (ORG). Using the transducer formalism, the authors in [17] proposed a system recognizing ANEs for the sport field. The proposed system is based on syntactic patterns transformed into transducers. Generally, the transducers are called in a specific passage order. This order is known as transducer cascade. Moreover, the transducer cascade is used for the Arabic language to recognize and annotate the ANEs for the Date [9], Person name [10] and Event [11] categories. In fact, the proposed transducer cascades were tested on Arabic Wikipedia corpus and generated using CasSys tool integrated under the linguistic platform Unitex [23].

The statistical approach. The statistical approach takes advantage of ML-algorithms to learn NER decisions from annotated corpora. The current approach requires the availability of large annotated data. Using this approach, we quote the system elaborated in [1] to recognize the ANEs, which can be assigned to 10 identified categories. The elaborated system includes the CRF and bootstrapping pattern recognition. In [34], the authors developed a system to recognize Arabic temporal expressions by exploiting the dashtag–TMP, which is a temporal modifier referring to a point in time or a time span. Furthermore, [26] developed a system recognizing four ANE categories adopted the Artificial Neural Networks (ANN). To improve the NER on microblogs, [14] proposed a system recognizing ANE using the Condition Random Field (CRF) classifier to tag new training set from Tweeter. In [39], the authors proposed a Biomedical NER (Bio-NER) based on a deep neural network architecture. This proposed architecture is composed of multiple layers. Indeed, each layer has abstract features based up on those generated by the lower layers. The elaborated method is absed on a Convolutional Neural Network (CNN) technique. The exploited unlabeled corpora are GENIA corpus and a set of data collected from PUBMED database. The collection from the PUBMED is made through the biopython[1] tool. Based on PUBMED database, [21] proposed a system to extract the biomedical NEs. To realize the proposed system, the authors chose a subset of the relevant document from the PUPMED. Then, they treated the collected corpus through a preprocessing task including the tokenization and the stemming steps. Besides, the preprocessing is a part of the NER phase. In fact, there are other tasks belong to the NER phase, such as the syntactic annotation like the Part of Speech (PoS) tagging and noun phrase chunking. The semantic annotation is

[1] http://biopython.org/wiki/Biopython.

the core of the NER process, which is made in this system using lexical resources and ML based on NLP Model generation using the CRF. Moreover, the system outcome compared to the SVM (Super Vector Machine) algorithm shows that FS-CRF (Feature space based CRF) does better than the SVM.

The hybrid approach. The hybrid approach is the fusion of both the symbolic and statistical approaches, which are complementary. This approach helps to achieve a significant improvement of NER performance. Among the systems based on this approach, we quote: The automatic system developed by [35] to recognize the NEs having the category Event. For the Turkish language, [22] developed a system to recognize the Turkish NEs. For the Arabic language, [32] proposed a system capable of recognizing 11 categories that can represent an ANE. The NER progresses in several languages while it remains a challenge in Indian languages, such as the Assamese. In reality, the Assamese language suffered greatly from the lack of research effort as well as the appropriate textual resources. Besides, the available corpora have a quite small number compared to other languages. For this reason, there are some studied that were carried out to develop systems, which can perform the NER in Assamese texts [37]. In [36], the authors elaborated is the first hybrid system to realize the NER in Assamese. This system recognizes the Assamese NEs and it is capable to treat four categories, such as Person, Location, Organization and Miscellaneous. The authors used three main steps for the NER process. The ML approach is the first step involving both the CRF and HMM techniques. The second step concerns the rule-based approach, which consists in using a set of handcrafted rules. The last step is the gazetteers-based approach, which integrates the NE tagging using lists including location, person and organization names. According to specific domains, [29] elaborated a hybrid model for NER in unstructured biomedical texts. The elaborated model is a framework that has a main task consisting of the identification and classification of the biomedical NEs into five classes namely DNA, RNA, protein, cell-in and cell-type. In fact, the proposed model combines the rule-based and ML approaches applied after a preprocessing step. The rule-based approach is dedicated to the NE identification. However, the NE classification is ensured by the ML approach, especially, the SVM classifier. The authors experimented their model on the data set from GENIA corpus (Medline abstract collection).

In the previous work, the authors adopted no formal definition determining the recognized NE limits. Furthermore, the illustrated NER systems are based on textual resources that are not always exhaustive. In fact, the free resources can be a solution for this difficulty. The identified rules for the NER can cause ambiguity problems if they are applied without a specific order. Therefore, this case requires an adequate formalism to ensure a good NER. Thus, using an annotation standard to detail NE component is necessary to produce structured corpora. In the following section, we will explain our linguistic study to identify ANEs from an Arabic Wikipedia corpus.

3 ANE Identification and Categorization

The objective of our study on the corpus extracted from Arabic Wikipedia is to have a formal and generic system applicable to all domains. This system allows the ANE recognition and treats the various ANE forms appearing in Wikipedia articles with a prediction of those that can be recognized. Generally, the article content is written in Modern Standard Arabic (MSA), Classical Arabic (CA) and Dialectal Arabic (DA) [4]. Anyway, this diversity of the language styles is very motivating and it helps us to collect numerous alternative forms. Indeed, it is not easy to identify an ANE because we must refer to a clear definition that helps determine its limits. Even, we should mention that the ANE categorization is not a trivial task. We know that there are various opinions about which categories should be regarded as an ANE and how the limits of those categories should be. If we want to detect a relevant ANE, then we must analyze all its presented forms while respecting several regional writings. Besides, we should unify the similar forms to eliminate the repeated ones and separate the different forms to avoid ambiguity. Furthermore, we determine an ANE as an expression that can or cannot contain a proper name. The determined ANE can be structured through categories and subcategories. The following sub-section describes our linguistic study to identify ANE categories and forms in the Arabic Wikipedia study corpus.

3.1 Identification of the ANE Forms and Categories

Our linguistic study shows that the Arabic Wikipedia study corpus comprises five main ANE categories, which are Date, Person name, Location, Event and Organization. Each category is composed of refined sub-categories.

We identify the "Date" category as a specific part of the numerical expressions. In fact, we find several forms that can describe this category. Moreover, These forms are divided into six sets, which are Period, Century, Year, Date based on Month, full Date or Season followed by a Year. These sets are presented in the following figure associated with the adequate examples.

Figure 1 shows the six sets regrouping the different identified forms, which describe the Date category. For example, we find that period the form can be calculated based on the month, the day, the century and the year. In addition, our study shows that some trigger words can appear in different morphological forms (plural, dual), such as "سنتي/ between" used to calculated period between 2 years.

The "Person name" category represents various forms of an Arabic person name. These different forms are highly associated with the country origin, religion, culture, level of formality and personal preference. Generally, this category contains the following parts, "ism", "kunya", "nasab", "laqab", and "nisba" [33, 11]. Therefore, we remark that a person name form regroups at least one of the already mentioned parts. For example, the ANE "عبد الفتاح أبو غدة/ Abd Al-fattah Abu Ghoda" is composed of an "ism", which is "عبد الفتاح/ Abd Al-fattah" and a laqab, which is

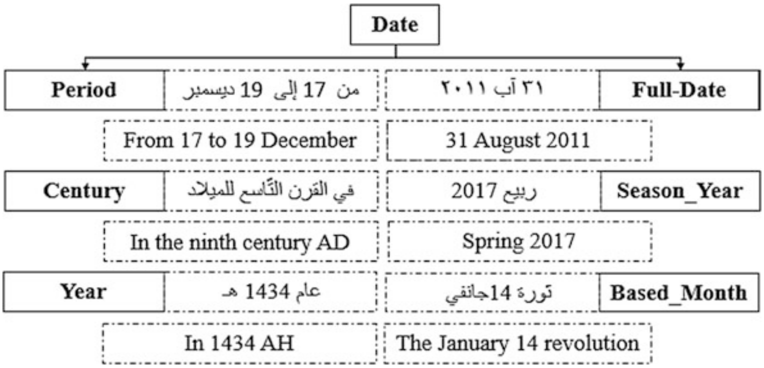


Fig. 1 Forms describing ANE date

“أبو غدة/ Abu Ghoda” expressing a kunya. Otherwise, we classify the trigger words preceding the ANEs in eight classes that are Civilities (الآنسة/ Miss), Profession (المدير/ the director), Peerage function (الأمير/ the prince), Political function (الوزير/ the minister), Religious function (المؤذن/ the muezzin), Sportive function (اللاعب/ the player), Artistic function (الممثل/ the actor) and Military function (الرائد/ the major).

The place names are the most common forms appearing in our study corpus. For this reason, we affect them to the category named “Location”. Then, we extend this category to be composed of three sub-categories, which are Absolute, Relative and Geographic Location. We mean by Absolute Location category all the place names defined in its sense by one place, such as the country (تونس/ Tunisia), continent (أفريقيا/ Africa), city (تور/ Tours), delegation (مناخة/ Manakhah) and region (الشبطلية/ Al-Shabtiliyah) names. The Absolute Location forms can be identified through the trigger words and prepositions playing the role of a place indicator. Moreover, we identify the Relative Location sub-category as a place name, which can be quantified by its relationship with an absolute Location as building. We identified 16 sub-categories describing this sub-category. The Relative Location forms are identified through trigger words, which can be a part of the identified ANEs. Grammatically, the trigger words can be defined or undefined as “المدرسة/ the school” and “مدرسة/ school”. The richness of our study in terms of Relative location sub-category enables us to refine it into 16 sub-categories. Then, we associate each sub-category with its appropriate forms.

Our identification of the Geographic Location sub-category considers them as specific physical points on earth. In addition, we notice that all the ANE forms related to this sub-category appear with geographic features, which can be a Mountain or a Hydronym. Similarly, the Hydronym forms are divided into two sets where the first set regroups the river or the lake names and the second set is dedicated to the sea names. We illustrate a small hierarchy describing the sub-categories of Relative and Geographic Location.

Figure 2 regroups the sub-categories composing both the Relative and Geographic Location. The forms describing the Relative Location sub-categories have

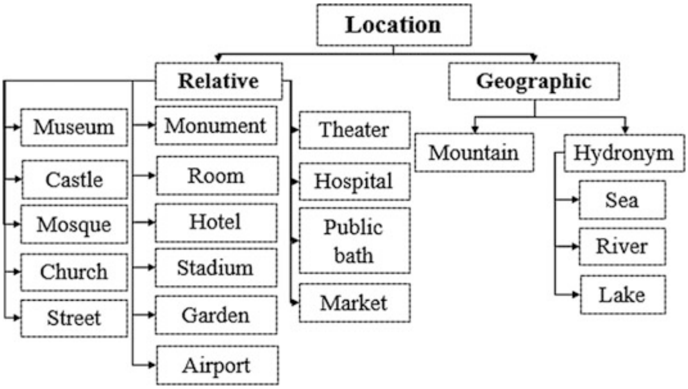


Fig. 2 Sub-categorization of relative and geographic location

different components. These components can be adjectives (الدولي/ the international), such as “المسرح/ الدولي the international theater”. It can be noun phrases (العلم الأردني/ the Jordanian flag) in monument names, such as “سارية العلم الأردني/ The Jordanian flagpole”. Then, the Relative Location sub-category can contain defined common noun (الثقافة/ culture), such as “شارع الثقافة/ Cultural street”. Similarly it can contains the Absolute Location (عمان/ Amman), such as “مطار عمان المدني// Amman civil airport”. The Geographic Location forms are also expressed through several trigger words, which help determine their features. For example, the trigger word “جبل/ mountain” is a mountain feature among a set of synonym trigger words, such as {جبل, جبال, مونتي, الجبل, تل, الجبال, قمة}. Notice that some Trigger words are foreign as “مونتي/ mountain” referring to other languages.

Besides, we identified the category Organization and its forms using a set of trigger words allowing not only their identification but also the determination of the organization nature. The organization names can describe an institution, a minister, a university, a society and so on. There are organizations that appear in the acronym form. Nevertheless, we do not treat this case because it is relatively rare in our study corpus.

Figure 3 describes the different sub-categories composing the organization ANEs. The identified forms enable us to treat the agglutination when two common nouns are related by a conjunction, such as “التربية والتعليم/ education and teaching” and analyze the adjectival phrase, such as “التلفزة الوطنية التونسية/ the Tunisian national television”, which contains a succession of defined adjectives where “التونسية” is an Arabic adjective expressing a “Nisba”.

The last category that we identified is Event, which was well studied in [11]. Therefore, we identify an event as a nominal composition that can have different forms and we classify its forms using the sub-categories. We use the same sub-categories of the ANE event that can be political, cultural and religious.

Figure 4 describes the sub-categories that we used to classify the ANEs having the Event category. Similarly, Event is determined through a set of significant trigger words, which are a part of the identified ANEs. After the ANE form

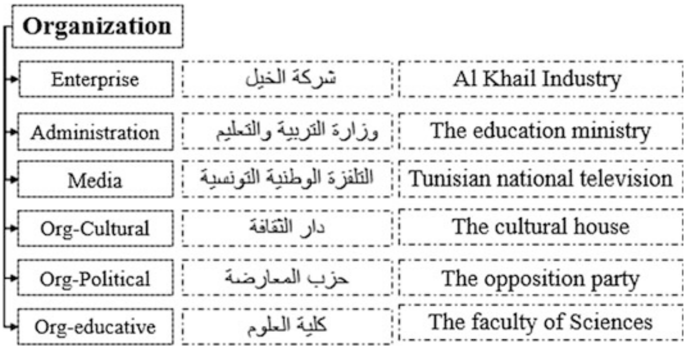


Fig. 3 Organization sub-categories

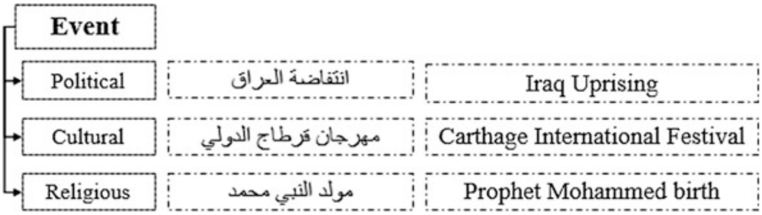


Fig. 4 Event sub-categories

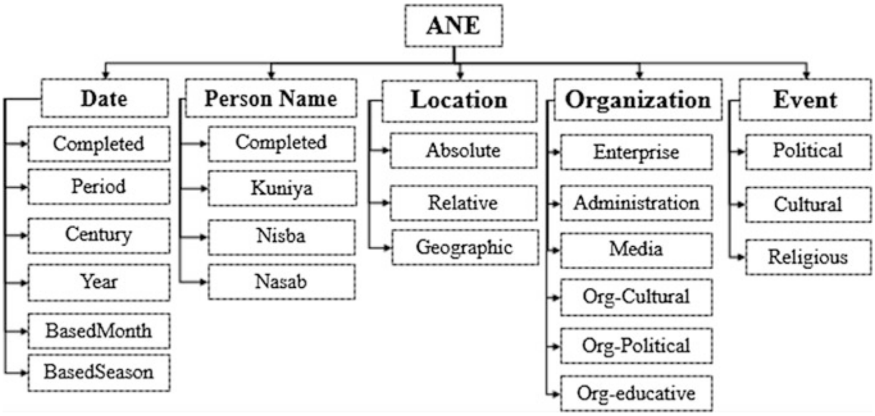


Fig. 5 ANE hierarchy identified from the study corpus

identification, we notice that the ANEs can be assigned to several categories and sub-categories. For this reason, we regroup them in an ANE hierarchy, which helps us to describe our categorization.

Figure 5 describes the ANE hierarchy associated with our study corpus. All the categories are refined to be composed of sub-categories. The illustrated categories are extended to increase the granularity level. This extension depends, in a large part, on the appearance of the ANE forms.

Our linguistic study plays a significant role not only in identifying the ANE forms but also in showing the different relationships that can relate them. In the following sub-section, we describe the different kinds of the relationships between the identified ANEs appearing in our Arabic Wikipedia study corpus.

3.2 Relationship Between ANEs

The relationship between the ANEs can be binary (involving two ANEs) or more complex to be an imbrication of ANEs. The ANE imbrication always describes a composition of these ANE without a specific link. However, when a particular link appears through phrases or prepositions, the kind of the relation becomes an SR. In what follows, we will describe the ANE imbrication.

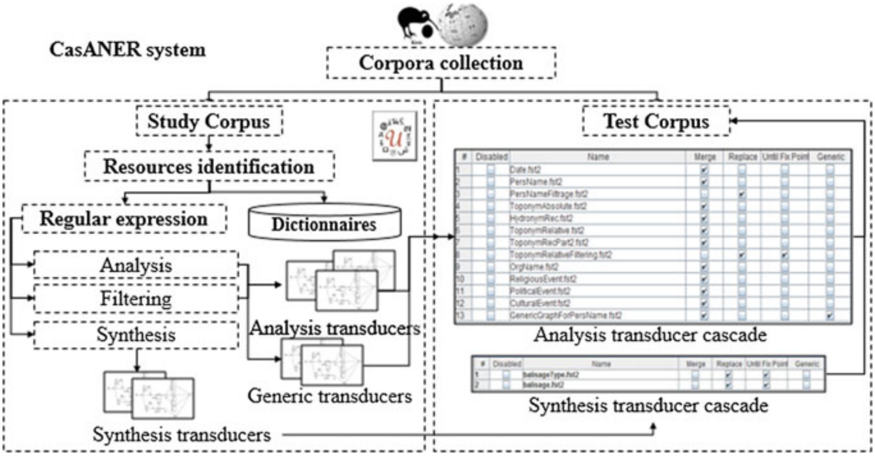
The ANEs having respectively the category Event and Location have a composition relationship with those having the Date category. For example, a relative location name, especially stadium one, can contain a relative date, such as “14جانفي ملعب/ 14 January stadium”. We notice that the date is composed of a form among those identified during our linguistic study. The imbrication of these ANEs may be surrounded by a left context, such as “ملعب 14جانفي بنابل/ 14 January stadium of Nabeul” and ملعب بالكاف/14 “جانفي January stadium of Kef”. The illustrated left context is in its turn an agglutinated absolute location name. Sometimes, the ANEs Date refer to symbolic events occurred in the past. For example, “يوم 14جانفي / 14th January” is a relative date referring to the event of the Tunisian revolution. Similarly, Event and Location can be a part of the Person Name category, such as “ملعب الطيب المهيري بصفاقس/ El-Taieb Mhiri stadium of Sfax” where “الطيب المهيري/ El-Taieb Mhiri” is a person name of a famous personality. A person name can also be integrated in an organization name, such as “مؤسسة العنود الخيرية/ Anoud Charitable Foundation” and we notice that here the name person is a princess first name. Our study corpus has a rich content, which is very interesting. This richness helps us to value the significant SRs between ANEs. The SRs are always binary relating the ANEs having the same or different categories. The same ANE categories can be related by a synonymy, such as “الثورة التونسية/ Tunisian revolution” and “ثورة الحرية والكرامة/ the freedom and dignity revolution”. The synonyms of ANEs describe an event that can be expressed by an ANE Date, which is “جانفي 14/ 14th January”. Among the SRs linking different categories, we find the meronymy, which expresses an inclusion relation, such as “كمال الملاخ/ Kamal Al-Mallakh” and “القاهرة/ Cairo”. This means that the first ANE is included in the second one. Generally, the linked ANEs surround the relevant expressions or prepositions, which facilitates to guess the SR type. In fact, determining SRs between ANEs is an important challenge that will be studied later. In the following section, we will describe our proposed method to recognize and annotate ANEs.

4 Proposed Method

Our proposed method, which is intended to elaborate CasANER system recognizing and annotating ANEs, is composed of the following steps: the collection of Arabic Wikipedia articles to construct corpora (study and test), the construction of necessary dictionaries and extraction rules and the transducer establishment. We propose the following architecture to describe the different steps.

Figure 6 shows that the CasANER system relies on an important step, which is the resource identification. Furthermore, we construct our extraction rules based on the trigger words, which are identified during the linguistic study. Besides, we convert these established extraction rules into regular expressions and we regroup them into three sets; analysis, filtering and synthesis. Hence, we distinguish two main phases; Analysis and Synthesis.

It should be recalled that a transducer is a graphic representation of the regular expression specification. This establishment is not arbitrary but we fix a principle to create its boxes. Therefore, each transducer contains the collected trigger words, particularly those having common paths. Besides, we always try to separate the overlapped paths because they may create ambiguities. Inside the created transducer, we can call sub-graphs that enables us to reduce the transducer size. According to the annotation, the used tags always encompass the recognized paths. In this context, the analysis and filtering transducers ensure the analysis phase. In fact, the analysis transducers deal with both the recognition and annotation processes. However, we construct the filtering transducers to recognize the ANEs that are not treated by the first analysis. These ANEs are not treated due to the preprocessing, especially, the segmentation step. Otherwise, we try to rectify the recognition paths using variables to organize the output of the filtering transducer. In what follows, we will describe the analysis transducers.



#	Disabled	Name	Merge	Replace	Until Fix Point	Generic
1	□	Date.fst2	☑	□	□	□
2	□	PersName.fst2	☑	□	□	□
3	□	PersNameFiltrage.fst2	□	☑	□	□
4	□	ToponymAbsolute.fst2	☑	□	□	□
5	□	HydronymRec.fst2	☑	□	□	□
6	□	ToponymRelative.fst2	☑	□	□	□
7	□	ToponymRecPart2.fst2	☑	□	□	□
8	□	ToponymRelativeFiltering.fst2	□	☑	☑	□
9	□	OrgName.fst2	☑	□	□	□
10	□	ReligiousEvent.fst2	☑	□	□	□
11	□	PoliticalEvent.fst2	☑	□	□	□
12	□	CulturalEvent.fst2	☑	□	□	□
13	□	GenericGraphForPersName.fst2	☑	□	□	☑

#	Disabled	Name	Merge	Replace	Until Fix Point	Generic
1	□	balisageType.fst2	□	☑	☑	□
2	□	balisage.fst2	□	☑	☑	□

Fig. 6 CasANER architecture

4.1 Analysis Transducer Establishment

The analysis transducers regroup recognition and annotation paths for the identified categories and sub-categories with several kinds of boxes. To analyze the morphological level, we use the morphological mode or filter, which facilitate the agglutination treatment. The morphological filter is used to control the form of a word belonging to the syntactic pattern, which recognizes an ANE. Moreover, the morphological mode is used, for example, to read a conjunction (<CnjCrd>) linked to a first name, like "ومحمد/ and Mohammed". In what follows, we will illustrate some transducers to show their specificity and form.

For the Date category, we create a main transducer, which regroups other transducers recognizing the identified forms (Sect. 3.1). Among these transducers, we illustrate the transducer that recognizes a season followed by a year.

Figure 7 shows the alternative paths to recognize a season followed by a year forms. This transducer resolves the case of the agglutination phenomena. This case is described by two boxes of "<" and ">" to mean that the recognition touches the morphological level. Therefore, the preposition (<Prps>) of the conjunction (<CnjCrd>) will be separated from the box containing <Np + season>. This transducer can recognize the following ANEs "2000 صيف/ summer 2000", "م 1990 عام ربيع/ spring 1990" when the recognition path detects the specific and internal indicators and "فصل الخريف/ the autumn season" if the ANE is detected through a trigger word belonging to the first box.

It is very important to notice that our analysis transducer annotated the recognized ANEs using {} markers. In fact, we do not use them arbitrarily because they make the recognized ANE polylexcial word, which cannot be detected by another transducer. This means that our recognized ANEs are protected. We should also notice that the annotation markers are represented in the node output. Besides, we use the same principle to treat the rest of the categories.

Regarding the Person name category, there are two main transducers. The first one calls the sub-transducers treating this category without trigger words. Nevertheless, there are sub-transducers that need to be preceded by a box storing the sub-graphs containing the trigger words, which are organized based on the

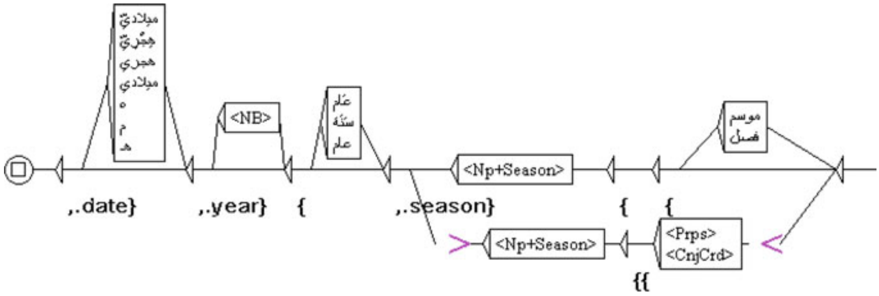


Fig. 7 Transducer recognizing ANE date composed of a season followed by a year

identified classes (Sect. 3.1). In addition, the use of trigger words helps avoid the ambiguity problems related to this category. However, we can decrease this ambiguity using a specific kind of transducer that will be explained later.

Figure 8 describes a path recognizing an identified form, which is composed of a first name, a civility and a last name. The illustrated path can recognize the ANE "العادلي فؤاد بك/ Foued bek Al-Adly", which contains a civility belonging the list stored in the sub-graph "Civilities". This transducer does not comprise the final tag "persName" because it will be called in the main transducer.

In the previous transducers, we have not use any element inside the tags because we have treated only the ANE forms related to a category. However, we will use an element named "type" to store the sub-category values. We begin by describing a transducer, which recognizes the Absolute Location sub-category to illustrate the element utility.

Figure 9 describes the recognition of the city names, which will be annotated through the "placeName + type = "city"" tag. It is worth noting that we use the "+" sign just to replace the space, which may cause ambiguity problems during the experimentation phase. Here, the element "type" helps determine the value of the Absolute Location sub-category. Besides, it can take "country", "continent", "delegation" and "region" depending on the Absolute Location nature. In the agglutination case, we add the "Prps" tag as output to the boxes containing "Prps" and "CnjCrd" in order to protect them. In fact, this absolute Location can be a part of other ANE. Thus, if we do not separate the ANE and the preposition by tags, then this absolute location will not be recognized.

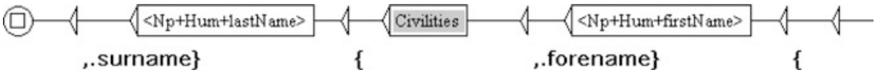


Fig. 8 Transducer recognizing an ANE having person name category

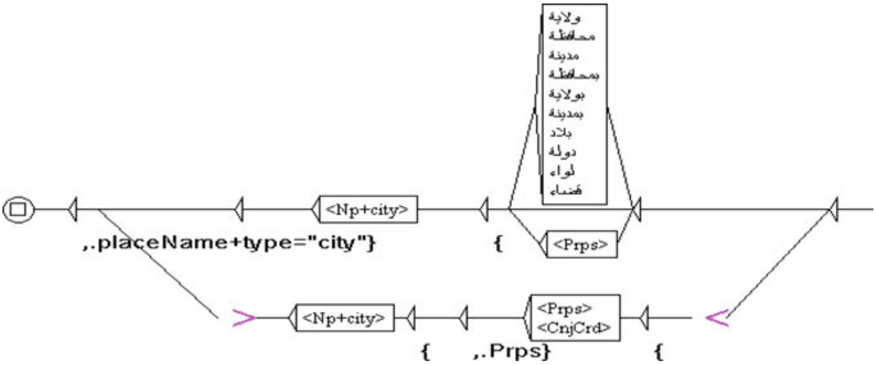


Fig. 9 Transducer recognizing city names

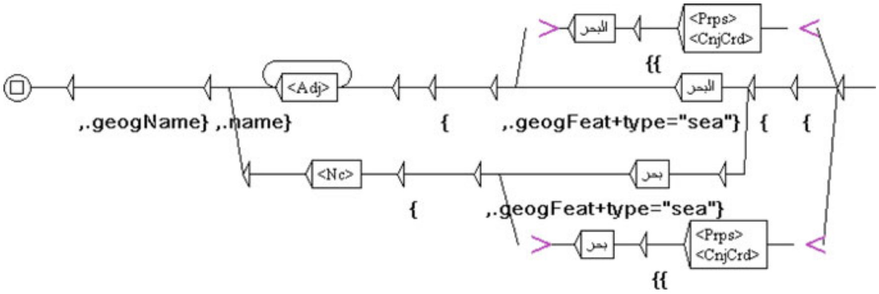


Fig. 10 Transducer recognizing a sea name

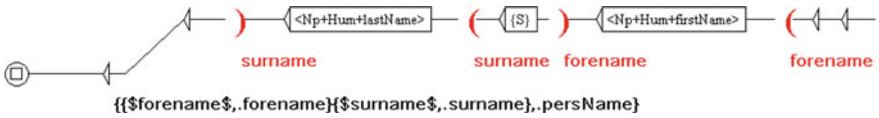


Fig. 11 Filtering transducer to recognize a person name

The "type" element is mainly used to describe a sub-category. Nevertheless, we can use it to describe the value of a geographic feature when we recognize the ANEs having the Geographic and Hydronym sub categories.

Figure 10 shows our manner to annotate the Hydronym sub-category, especially the sea names. At the begging of each path, we consider the trigger words as geographic features associated with a tag called "geogFeat". Then, we surround the rest of the ANEs with a tag entitled "geogName". The illustrated transducer is called in the main one where these paths will be integrated in a global tag "placeName" containing an element type = "Hydronym".

Filtering analysis transducer. The filtering transducers belong to the analysis phase. Their objective is to recognize the ANEs when their components are separated during the segmentation phase. It should be noted that the segmentation is made through a graph provided by the exploited linguistic platform. The filtering transducer is based on the same paths of the analysis one but each box is surrounded by variables to temporary store its value. At the end of the recognition path, we organize the output to obtain an annotated ANE.

Figure 11 describes a transducer treating an exceptional case for the Person name recognition. Here, we have two ANE components separated by {S}, which is a segmentation symbol. The illustrated transducer respects the same annotation principle of the analysis transducer. In addition, we call the used variables (forename and surname) using the $ symbol to retrieve their values.

Generic transducer. Our generic transducers are tagging generalization graphs that aim to locate unrecognized NE occurrences when these NEs appear out of the context with those identified based on specific contexts [28]. Creating a tagging generalization graph consists in building a path that begins with a box having $G

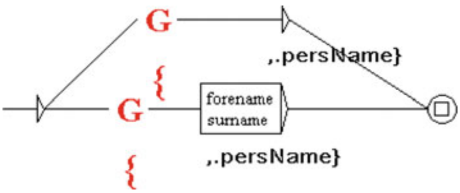


Fig. 12 Tagging generalization graph for the category person name

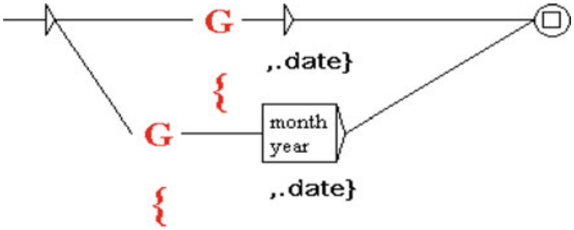


Fig. 13 Tagging generalization graph for the category date

and an opening curly bracket output. Then, the same path has a second box containing the searched categories to find the unrecognized NEs. Checking the generalization mode is necessary when we add this transducer kind to our cascade using the exploited linguistic platform. This enables to consult a text file named "tok_by_alph.txt" of the previous graph placed in the same transducer cascade. In reality, the tagging generalization graph stores in the box all the NEs of the searched category, which are extracted from the "tok_by_alph.txt" file. Finally, the graph recognizes the NEs out of context and attributes the main categories. For example, if the graph recognizes a forename from a full-recognized ANE, then, it will assign it to the "persName" category.

Inside the tagging generalization graph, we can put some restrictions in the second box containing the searched category as an output node. This case is used to treat the recognized ANE components. In fact, we can add new information to the category described in the output node to complete the annotation. The following figure illustrates the restriction created to recognize only forenames and surnames out of the context.

Figure 12 illustrates a tagging generalization graph treating the Person name category. In fact, we duplicate the box containing $G for each path. Moreover, the restriction here is described in the second path.

We also use the restriction principle to recognize the elements composing the Date forms, which helps us to recognize the months and years out of the context (Fig. 13).

The transducers, ensuring the analysis phase, contain an annotation form respecting the tool, which will regroup them into a cascade. However, we organize the annotation syntax as a preprocessing to transform it into the TEI annotation. In what follows, we will describe the principle of this transformation using the synthesis transducers.

4.2 Synthesis Transducer Establishment

The synthesis transducer consists in transforming the annotation made by the analysis phase to the annotation related to the TEI recommendation. It is worth noting that the TEI recommends an international consortium where their goals are the development of a set of standards for the preparation and the exchange of electronic texts [7, 24]. Therefore, we establish our synthesis transducers using the syntax described in [38].

The TEI syntax is defined as follows: an opening tag describing such category, like <persName> and a closing tag </persName> that surround the ANE. The tag <persName> can include an imbrication of the first name, last name and trigger word that can precede a person name that are surrounded respectively by other tags, such as <forename>, <surname> and <roleName>. In the "roleName" tag, it is possible to specify the roleName type, such as military function. Moreover, the addition of the sub-category is made using an element named "type", which can contain different values. Indeed, other categories of an ANE can be presented by the TEI recommendation, such as <orgName> and </orgName> to describe an organization names and <placeName> and </placeName> to describe the Location category, which can have an element "type", such as type = "castle" to annotate castle names, which are Relative Location. To understand the principle of the transformation of the analysis annotation into the TEI recommendation tags, we propose the following architecture (Fig. 14).

After carrying out an analytical study on the annotated files, we have developed two transducers for the synthesis phase. The first one treats the recognized ANEs, which do not have a sub-category. Based on the linguistic study and the refined categorization, we elaborate a second transducer to transform the annotation of the ANEs having an element called "type", including their sub-categories.

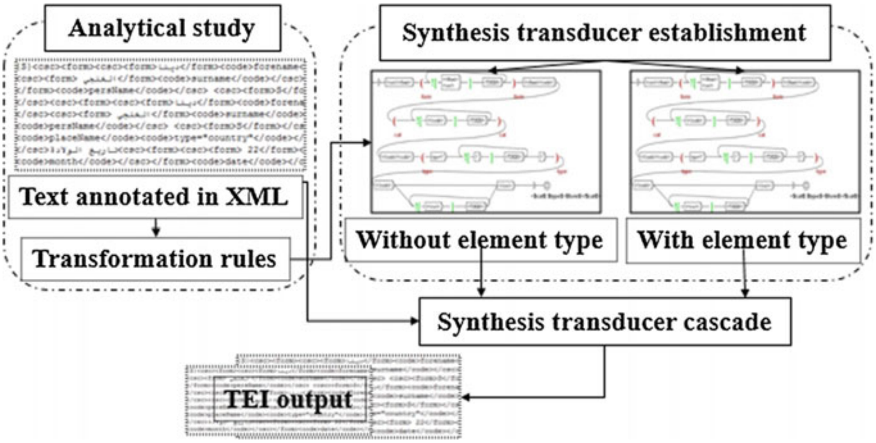


Fig. 14 Principle of transforming annotation

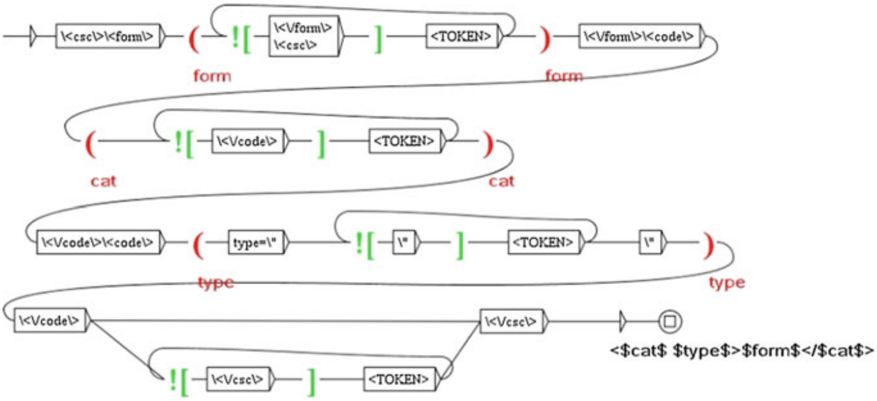


Fig. 15 Transducer transforming the annotation of the analysis phase into TEI

Figure 15 describes the transducer transforming the annotation specific to the analysis phase into TEI when the recognized ANE contains an element “type”. In fact, we propose a path taking as input the XML (eXtensible Markup Language) format of the annotated text files generated by the tool creating the transducer cascade. Hence, we use the negative context through ![,] markers and variables using (,) markers, which organize the output annotation.

5 Implementation and Evaluation

Our CasANER system is implemented using the linguistic platform Unitex[2] (version 3.2 alpha), especially, the CasSys tool to generate transducer cascades. However, we experimented our proposed system through a test corpus, which is also collected from Arabic Wikipedia with Kiwix[3] tool. The collected test corpus contains text files for a cumulative of 95 378 tokens. Furthermore, the study corpus, containing text files for a cumulative of 146 000 tokens, enables us to create new dictionaries and update dictionaries available under Unitex platform.

In some cases, we find that Arabic dictionaries under Unitex elaborated by [16] do not respond to our needs in order to make a good recognition. For this reason, we exploit Arabic Wikipedia, precisely our study corpus to improve their coverage or to create new dictionaries. In fact, creating a new dictionary requires respecting the features proposed in a tagset[4] available under Unitex. Indeed, they store different variations that an Arabic entry can have.

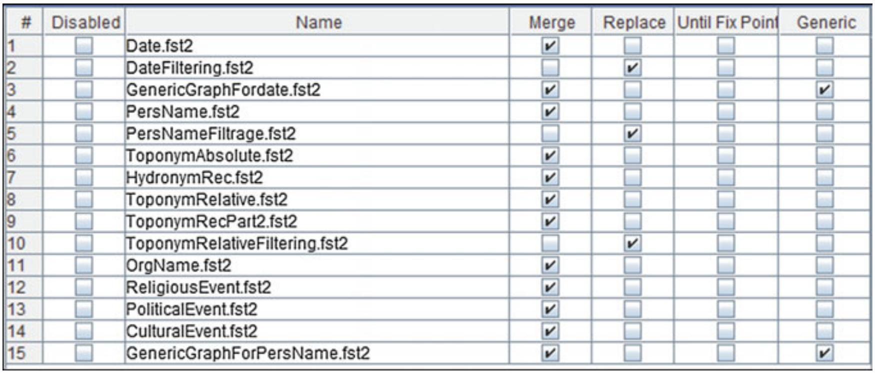
[2] http://www-igm.univ-mlv.fr/∼unitex/.

[3] http://wiki.kiwix.org/wiki/Main_Page.

[4] The tagset of Arabic Unitex package dictionaries.

Table 1 Dictionary coverage

Dictionary name	Coverage
First name	9 917
Last name	1 991
Arabic adjective	2 938
Toponym	13 757
Common noun	14 976
Direction	14
Season	11
Day	23
Month	48



#	Disabled	Name	Merge	Replace	Until Fix Point	Generic
1	☐	Date.fst2	☑	☐	☐	☐
2	☐	DateFiltering.fst2	☐	☑	☐	☐
3	☐	GenericGraphFordate.fst2	☑	☐	☐	☑
4	☐	PersName.fst2	☑	☐	☐	☐
5	☐	PersNameFiltrage.fst2	☐	☑	☐	☐
6	☐	ToponymAbsolute.fst2	☑	☐	☐	☐
7	☐	HydronymRec.fst2	☑	☐	☐	☐
8	☐	ToponymRelative.fst2	☑	☐	☐	☐
9	☐	ToponymRecPart2.fst2	☑	☐	☐	☐
10	☐	ToponymRelativeFiltering.fst2	☐	☑	☐	☐
11	☐	OrgName.fst2	☑	☐	☐	☐
12	☐	ReligiousEvent.fst2	☑	☐	☐	☐
13	☐	PoliticalEvent.fst2	☑	☐	☐	☐
14	☐	CulturalEvent.fst2	☑	☐	☐	☐
15	☐	GenericGraphForPersName.fst2	☑	☐	☐	☑

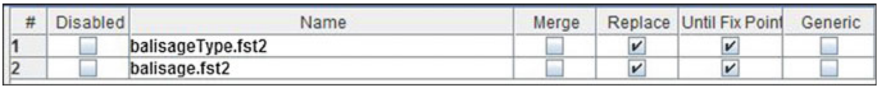
Fig. 16 First transducer cascade composing the CasANER system

Table 1 shows the coverage of our dictionaries. In fact, to increase the coverage of the Arabic Adjective and Common noun dictionaries, we improve them using an automatic enrichment based on textual resources.

Our elaborated transducer cascade for the analysis phase calls 15 main graphs in a specific order and respects the adequate mode for each graph. However, we created 178 graphs including analysis, filtering and generic one.

Figure 16 shows the passage order that we have chosen to organize our analysis transducer cascade. In fact, there are graphs that are applied on the text using the mode “Merge”. However, the filtering graphs use the “Replace” mode because they use variables at the end of their recognition paths. Therefore, they need to replace the last ANEs with new ones associated with an annotation. The last use mode is Generic for applying the tagging generalization graph.

The generation of the final analysis transducer cascade is not an easy task. In reality, it needs to change the transducer order to fix the best one and test it. During this test task, we detected several kinds of errors. For example, when we treat the category Event we must separate the graphs depending on these sub-categories. Then, if we put cultural event graph before the religious one. We found that the



#	Disabled	Name	Merge	Replace	Until Fix Poin	Generic
1	□	balisageType.fst2	□	☑	☑	□
2	□	balisage.fst2	□	☑	☑	□

Fig. 17 Form of the synthesis transducer cascade

ANE "مهرجان عيد الفطر/ Eid Al-fitr festival" was not recognized because it contains the ANE "عيد الفطر/ Eid Al-fitr" which is a religious festival.

The second part composing the CasANER system is the synthesis transducer cascade. This cascade is dedicated to deliver a structure output by transforming CasSys tags to TEI tags. The synthesis cascade also needs a passage order but it is not difficult to choose the best order. Furthermore, it differs from the analysis cascade in the passage mode and in the format of the input file. In fact, the difference of the used passage mode is justified by the fact that the transducers called by this cascade use variables to organize the output. For this reason, the last ANE will be replaced by the same ANE with the new tags.

Figure 17 shows the passage order of our synthesis transducer cascade. In other words, we pass the graph treating the recognized ANEs having an element. The new mode here is "Until fix point", which means that this graph is applied once or re-applied several times until no change occurs in the text. The role of this synthesis transducer is summarized as follows: initially the ANE is annotated using { }, which is a representation specific to CasSys tool. The accolade annotation has its appropriate translation described in XML, which uses significant tags, such as <csc> means the ANE annotated through the cascade, <form> containing the ANE value and <code> detailing the category in which an ANE is assigned. This tag contains other <code> to describe the sub-category of an ANE using the element type. Besides, the TEI recommendation tags replace the already mentioned ones.

Applying our CasANER system, including the two kinds of transducer cascades, provides a structure and normalized corpus (extracted initially from Arabic Wikipedia). This structure output can be used by several NLP-application. However, before its exploitation, we must evaluate the CasANER performance to show its efficiency.

It should be recalled that evaluating our proposed CasANER system is an important process that helps us to prove its reliability. For this reason, we evaluated it in two manners. The first manner is by calculating the measure values and the second phase by applying CasANER on a new corpus annotated by a ML based system. The second evaluation also permits comparing the result of both systems based on different approaches.

The first CasANER evaluation is performed by the precision, recall and F-score measures that are illustrated in Table 2. We should improve the analysis transducer cascade to cover all the ANEs.

Table 2 demonstrates that CasANER shows a precision of 92%, a recall of 91% and an f-score of 91% for ANE. Therefore, we find that the obtained outcomes are very motivating. Notice that the number of ANE detected in error causes the obtained recall value. The Errors presented in this recognition process are due to the

Table 2 CasANER evaluation using measure values

Recall	Precision	F-score
0.91	0.92	0.91

Table 3 CasANER evaluation by categories

	Date	Person name	Event	Location	Organization
Recall	0.78	0.87	0.92	0.95	0.99
Precision	0.81	0.95	0.96	0.94	0.97
F-score	0.79	0.90	0.93	0.94	0.97

fact that the dictionary coverage must be improved. In fact, the performance of the recognition increases if dictionary coverage is enriched. Errors can be caused by the structure of Arabic Wikipedia's articles. For example, there are ANEs having miss-spelled trigger words, such as "اتنفاضة الصدر" instead of "انتفاضة الصدر/ Sadr's uprising" and "محافة" replacing the word "محافظة/ city" determining city names. Furthermore, the errors can be found in the prepositions, which can play the indicator role to determine the ANE limits, such as "قي" instead of "في/ in".

In order to show the CasANER reliability, we propose an evaluation for each category using also measure values. This decomposition helps us to determine the parts that require an enhancement.

Table 3 proves that our CasANER system excels, especially, in the recognition of the Event, Location and Organization categories. The recognition of the Person name category is interesting because we used three kinds of transducers that are analysis, filter and generalized tagging, which help us to recognize several ANEs. The use of generalized tagging transducer permits to avoid ambiguity problems, for example, we must guess if a word is a first name or surname and it is not an adjective or a common noun. The Event, Location and Organization recognition relies on the important number of the identified extraction rules. However, the year recognition without trigger words and through prepositions cause the obtained measure values for the Date category.

The second evaluation consists in applying our proposed system on ANERcorp.[5] This corpus is freely distributed. [6] collected ANERcorp from several sources to obtain a generalized corpus. The ANERcorp was collected to construct the study and test corpora for the elaborated ANERsys. The ANERSys recognizes and annotates ANE using ML approach. ANERcorp contains more than 150 000 words annotated for the NER task. Each word in this corpus is annotated as one of the tags illustrated in Table 4.

The evaluation admits three steps. The first one consists in deleting tags existing in ANERcorp to recover the initial corpus. The second step is the application of our analysis transducer cascade on the initial ANERcorp to provide an annotated one.

[5] http://users.dsic.upv.es/grupos/nle/?file=kop4.php.

Table 4 Tags used by ANERsys

Tag	Signification
B-PERS	Beginning of the name of a PERSon
I-PERS	Inside of the name of a PERSon
B-LOC	Beginning of the name of a LOCation
I-LOC	Inside of the name of a LOCation
B-ORG	Beginning of the name of an ORGanization
I-ORG	Inside of the name of an ORGanization
B-MISC	Beginning of an NE that does not belong to any previous class called MISCellaneous
I-MSC	Inside of an NE that does not belong to any previous class
O	Annotated word is not an NE (Other)

The recognized ANEs inside the new corpus are annotated using { }, for this reason, it will be the input of the synthesis transducer cascade. We added a new transducer that is dedicated to transform the Location, Person and Organization tags into tags having the following format: <category> and </category> as <LOC> and </LOC> to replace <placeName> and </placeName>. The categories inside the new illustrated tags are the same used by the ANERsys. Thus, the annotated ANERcorp must be adopted to the same format generated by the synthesis transducer cascade. In the third step, the transformation transducer cascade ensures this adoption.

The transformation transducer cascade acts on the ANERCorp to transform the used tags. The first called transducer is dedicated to delete the Inside tags (i,e. I-LOC, I-PERS) and to regroup the annotated words with the beginning of the name as B-LOC in the same line. The second transducer takes the B-LOC, B-PERS, B-ORG and B-MISC and transforms these categories into the new format <Category> and </Category>. For example, the ANE “فرانكفورت/ Frankfurt” was annotated as [فرانكفورت B-LOC] and it is transformed to <LOC> فرانكفورت </LOC>. The last transducer deletes the O tag. All transducers are in mode “replace”. The two first transducers are applied in “until fix point” mode in order to treat imbrication inside the ANE. We also applied our synthesis cascade to facilitate the comparison. A new transducer is added to convert the TEI tags related to Location, Person and Organization to the already mentioned format <category> and </Category> tags.

The evaluation before the application of our analysis and synthesis transducer cascade on the ANERCorp is performed by the precision, recall and F-score measures that are illustrated in Table 3. Our measure values are presented to be compared with the ANERsys measure values that were tested on ANERCorp.

Table 5 illustrates the measure values related to our CasANER system and to ANERsys. We can see that the CasANER results are as efficient as ANERsys one in the recognition and annotation of PERS and ORG. However, the ANERsys can recognize and annotate ANE having the category LOC more than our system.

Table 5 Comparison between CasANER and ANERsys using measure values

	ANERSys			CasANER using ANERcorp		
	LOC	PERS	ORG	LOC	PERS	ORG
Recall	0.78	0.41	0.31	0.71	0.81	0.58
Precision	0.82	0.54	0.45	0.66	0.76	0.55
F-score	0.80	0.46	0.36	0.67	0.78	0.63

There are some cases undetected by CasANER, such as abbreviations. In fact, we do not treat this form, which frequently appears in ANERCorp especially in organization names such as <ORG> سي أن أن </ORG> and <ORG> في تي جي </ORG>. In fact, for the category Organization, we recognize it using trigger words. However, we do not treat the famous organization names as "الفيفا/ FIFA" (<ORG> ا لفيفا</ORG>). The NEs written in other languages are not treated by our proposed system, whereas there are several foreign NEs having the category PERS and ORG annotated in ANERCorp as <PERS> Charles I </PERS> and <ORG> El Telegramma Del Rif </ORG>. There are also ANE annotated in error by the ANERsys as "السودان" that was annotated as a person name.

The CasANER application contributes to the enrichement of the ANERCorp. This enrichment was made through, on the one hand, the refinement provided by the use of categories and sub-categories and, on the other hand, through the use of TEI tags. Moreover, our analysis transducer cascade, which is included in CasANER, can resolve the agglutination phenomena untreated in ANERCorp. For example, the ANE "وقاسم العزام/ and Quasim Alaazam" is annotated as follows <PERS> وقاسم العزام </PERS>. However, CasANER annotates this ANE as follows: the conjunction "و/ and" is surrounded by "Prps" tag as <Prps>و </Prps>. This Prps is attached to two sub-tags: <PERS> <forename> قاسم </forename> <surname> العز ام </surname> </PERS>.

6 Conclusion

In the present work, we proposed CasANER system to recognize and annotate ANEs. The system is composed of two kinds of transducer cascades, which are the analysis and the synthesis implemented through the linguistic platform Unitex, especially, using the CasSys tool.

To realize CasANER, we made a deep and detailed ANE categorization in order to develop a category hierarchy. The elaborated hierarchy relies on representative an Arabic Wikipedia corpus containing articles extracted from several Arabic countries. In fact, analysis cascade regroups transducers, which ensure the analysis, filter and generalization tagging phases. The filtering phase is intended to rectify the paths of the analysis transducers in order to have structured recognized ANEs. However, the generalization-tagging phase helps us to improve our system performance. According to the synthesis, this cascade regroups transducers that helps transform the annotation of the recognized ANEs into an annotation respecting the TEI recommendation. This transform enables us to generate structured output corpus that can be used by several NLP-applications. The evaluation using measure values shows that the CasANER proves its reliability because the obtained results are encouraging. Indeed, our system also demonstrates that it provides results more efficient than those of the ANERsys during the comparison process.

In future work, we will exploit our system output to extract the relevant SRs between the recognized and annotated ANEs in order to create electronic ANE dictionary having a convivial interface. In addition, we will improve the CasANER by adding a ML module to enhance its performance. Finally, we will focus on ameliorating our proposed system to realize the EL task based on existing free resources.

References

1. AbdelRahman, S., Elarnaoty, M., Magdy, M., Fahmy, A.: Integrated machine learning techniques for Arabic named entity recognition. Int. J. Comput. Sci. (IJCSI) 27–36 (2010)
2. Aboaoga, M., Aziz, M.J.A.: Arabic person names recognition by using a rule based approach. J. Comput. Sci. 922–927 (2013)
3. Aliane, H., Guendouzi, A., Mokrani, A.: Annotating Events, Time and Place Expressions in Arabic Texts. In: Proceedings of Recent Advances in Natural Language Processing, pp 25–31, Hissar, Bulgaria, 7–13 (2013)
4. Alsayadi, H.A., ElKorany, A.M.: Integrating semantic features for enhancing arabic named entity recognition. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(3), 2016 (2016)
5. Arnulphy, B., and Tannier, X.: Entités nommées événement: guide d'annotation. Notes ET Documents LMSI N: 2013–12 (2013)
6. Benajiba, Y., Rosso, P., Benedíruiz, J.M.: Anersys: An Arabic named entity recognition system based on maximum entropy. In: International Conference on Intelligent Text Processing and Computational Linguistics, pp. 143–153 (2007)
7. Ben Ismail, S., Maraoui, H., Haddar, K., Romary, L.: ALIF editor for generating Arabic normalized lexicons. In: Will Appear in Proceedings of the International Conference on Information and Communication Systems (ICICS) (2017)
8. Ben Mesmia, F., Friburger, N., Haddar, K., Maurel, D.: Construction d'une cascade de transducteurs pour la reconnaissance des dates à partir d'un corpus Wikipédia. Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, pp 8–11, Sousse, Tunisie (2015)
9. Ben Mesmia, F., Friburger, N., Haddar, K., Maurel, D.: Arabic named entity recognition process using transducer cascade and Arabic wikipedia. In: Proceedings of Recent Advances in Natural Language Processing, pp 48–54, Hissar, Bulgaria (2015)
10. Ben Mesmia, F., Friburger N., Haddar, K., Maurel, D.: Transducer cascade for an automatic recognition of Arabic Named Entities in order to establish links to free resources. In: First International Conference on Arabic Computational Linguistics (ACLing). pp 61–67 (2015)
11. Ben Mesmia, F., Friburger, N., Haddar, K., Maurel, D.: Recognition and TEI annotation of arabic event using transducers. In: Will appear in IEEE proceedings of CiLing'16 (2016)
12. Btoush, M.-H., Alarabeyyat, A., Olab, I.: Rule based approach for Arabic part of speech tagging and name entity recognition. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **7**(6), 331–335 (2016)

13. Chinchor, N.: Overview of MUC-7/MET-2. In Proceedings of the Seventh Message Understanding Conference (MUC-7), p 1–4, Fairfax, VA, USA (1998)
14. Darwish, K., Gao, W.: Simple effective microblog named entity recognition: Arabic as an example. In: LREC, pp 2513–2517 (2014)
15. Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R.: The automatic content extraction (ACE) program tasks, data, and evaluation. In: Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2004), pp. 837–840, Lisbon, Portugal (2004)
16. Doumi, N., Lehireche, A., Maurel, D., Ali Cherif, M.: La conception d'un jeu de ressources libres pour le TAL arabe sous Unitex. Paper presented at the TRADETAL2013, Colloque international en Traductologie et TAL, Oran—Algeria, 5–6 may. pp. 5–6 (2013)
17. Fehri, H., Haddar, K., Hamadou, A.B.: Recognition and translation of Arabic named entities with NooJ using a new representation model. In: Constant, M., Maletti, A., Savary, A. (eds.) FSMNLP, 9th International Workshop, pp. 134–142. ACL, Blois, France (2011)
18. Gravier, G., Bonastre, J.F., Galliano, S., Geoffrois, E., Mc Tait, K., Choukri, K.: ESTER, une campagne d'évaluation des systèmes d'indexation d'émissions radiophoniques, Proc. Journées d'Etude sur la Parole (2004)
19. Grouin, C., Rosset, S., Zweignbaum, P., Fort, K., Quintard, L.: Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview. In Proceedings of Linguistic Annotation Workshop, pp. 92–100 (2011)
20. Grishman, R., Sundheim, B.: Message understanding conference—6: a brief history. In: Proceedings of the 16th conference on Computational linguistics (COLING'96), pp 466–471, Copenhagen, Denmark (1996)
21. Kanya, N., Ravi, T.: Named Entity recognition from biomedical text—an information extraction task. ICTACT J. Sort Comput. **06**(04), 1302–1307 (2016)
22. Küçük, D., Yazici, A.: A hybrid named entity recognizer for Turkish. Expert Syst. Appl. **39** (3), 2733–2742 (2012)
23. Maurel, D., Friburger, N., Eshkol, I., Antoine. J.-Y.: Explorer des corpus à l'aide de CasSys. Application au Corpus d'Orléans. G. Willems (ed.). Texte et corpus n°4, Actes des 6es Journées Internationales de Linguistique de Corpus (JLC). pp 189–196 (2013)
24. Maraoui, H., Haddar, K.: Automatisation de l'encodage des lexiques arabes. Colloque pour les Étudiants Chercheurs en Traitement Automatique du Langage naturel et ses applications, pp 74–82, Sousse, Tunisie (2015)
25. Merchant, R., Okurowski, M., Chinchor, N.: The multilingual entity task (MET) overview. In: Proceedings of a workshop on held at Vienna, Virginia. Morristown, NJ, USA. Association for Computational Linguistics. pp 445–447 (1996)
26. Mohammed, N.F., Omar, N.: Arabic named entity recognition using artificial neural network. J. Comput. Sci. 1285–1293 (2012)
27. Oudah, M., Shaalan, K.: NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic. Nat. Lang. Eng. 1–32 (2016)
28. Paumier, S.: UNITEX 3.2 ALPHA. User Manuel. Université Paris-Est Marne-la-Vallée. Date of version, February 23, 2017. 383 p. (2017)
29. Ramesh, D., Sanampudi, S.-K. A Hybrid model for Named Entity Recognition in Biomedical text. Int. J. Sci. Eng. Res. **7**(6), 1164–1166. (2016). ISSN 2229-5518
30. Shaalan, K., Raza, H.: Person named entity recognition for Arabic. In: Proceedings of the 5th Workshop on Important Unresolved Matters, pp. 17–24 (2007)
31. Shaalan, K., Raza, H.: NERA: named entity recognition for Arabic. J. Am. Soc. Inform. Sci. Technol. **60**(9), 1652–1663 (2009)
32. Shaalan, K., Oudah, M.: A hybrid approach to Arabic named entity recognition. J. Inf. Sci. **40** (1), 67–87 (2014)
33. Shaalan, K.: A survey of Arabic named entity recognition and classification. Comput. Linguist. **40**(2), 469–510 (2014)

34. Saleh, I., Tounsi, L., Van Genabith, J.: ZamAn and Raqm: extracting temporal and numerical expressions. In Arabic in Information Retrieval, Lecture Notes in Computer Science, vol. 7097, pp. 562–573 (2011)
35. Serrano, L., Charnois, T., Brunessaux, S., Grilheres, B., Bouzid, M.: Combinaison d'approches pour l'extraction automatique d'événements. In: TALN'2012, volume 2, p: 423–430, Grenoble, France (2012)
36. Sharma, P., Sharma, U., Kalita, J.: Named Entity Recognition in Assamese: A hybrid approach. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI-2016), Jaipur, India (2016)
37. Sharma, P., Sharma, U., Kalita J.: Named entity recognition in assamese. J. Comput. Appl. **142**(8), 1–8 (2016)
38. Text Encoding Initiative Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL. Version 3.1.0. 1887 p. (2016)
39. Yao, L., Liu, H., Liu, Y., Li, X., Anwar, M.-W.: Biomedical named entity recognition based on deep neutral network. Int. J. Hybrid Inf. Technol. **8**(8), 279–288 (2015)