Arabic Language in the Context of Information Extraction Task

Meshrif Alruily, Aladdin Ayesh, Hussien Zedan
Software Technology Research Laboratory
De Montfort University
The Gateway, Leicester, LE1 9BH UK
{meshrif, aayesh, hzedan}@dmu.ac.uk

ABSTRACT: *In the past few years, researchers have started paying attention to the Arabic language. In this paper we review information extraction systems that were developed for the Arabic to extract predefined entities. A comparisons are conducted between these systems in terms of their performance in extracting the common entities, the approach used whether rule-based or machine learning and type of corpus.*

Keywords: Rule based corpus, Arabic language processing, Entity recognition systems, Machine learning, Arabic text nouns

Received: 1 March 2011, Revised 8 April 2011, Accepted 19 April 2011

1. Introduction

According to Fan et al. [1], the basic process in analysing textual data is information extraction, and it is particularly useful when dealing with vast volumes of text. The beginning for extracting specific types of information from particular domains was in the late 1980s; the Defence Advance Research Project Agency (DARPA) initiated a series of Message Understanding Conferences (MUC). In MUC-6 [2], [3], Named Entity Recognition (NER) was introduced. NER falls under the information extraction domain, and its mission is to identify entities from a text and to classify them into predefined categories, which include: personal names, locations, organizational names, dates, times, money and percentages [2]. So, the NER task can be divided into two sub-tasks: entity detection and entity classification [4]. Recently, two other disciplines related to NER have emerged: Automatic Content Extraction (ACE) and the Conference on Natural Language Learning (CoNLL). The former focuses on entity types, including person, organization, location, facility, weapon, vehicle and geopolitical entities, whereas, the latter concentrates on only four entity types: person, organization, location, and miscellaneous entities [5].

In this paper, information extraction approaches are described in section II. The Arabic information extraction systems use rule-based approach are examined in Section III. Section IV presents the Arabic information extraction systems based on machine learning approach. Finally, the conclusion is provided in section V.

2. Information Extraction Approaches

Several approaches have been proposed for performing information extraction tasks, they can be divided into three categories, as follows:

2.1 Hand-crafted rules, known as linguistic approaches.

This approach is usually used for extracting information from specific domains. It is simple to build because its goal is only to fill out templates, i.e. it is not document understanding. Moreover, it can be trained on annotated or unannotated corpora [6]. Although it can achieve reasonable levels of success, there are some disadvantages [7], [8]. For instance, it is time consuming because it is slow to build, and it is difficult to scale to new domains [8].

Toral [7] explained that this approach is applied using rules and gazetteers. The Gate system was developed at Sheffield University, and is a type of software that follows this approach. The task of this system is to extract named entities [9].

2.2 Machine learning approaches

Because of difficulties that faced researchers when building hand crafted rules, a need for automatically leaning extraction rules emerged [10]. There are three types of machine learning approach: supervised, semi-supervised and unsupervised. According to Nadeau [2], most developed systems have been designed based on handcrafted rule based systems or supervised learning based systems.In both approaches, a corpus must be studied and analysed by hand to gain sufficient knowledge to build the rules or to feed the machine learning algorithms. However, the supervised learning techniques, which include Hidden Markov Models (HMM), Maximum Entropy (ME), Support Vector Machines (SVM) and Conditional Random Fields (CRF) need a large annotated corpus for designing the systems. As a result, this disadvantage of supervised machine learning led to the emergence of semi-supervised and unsupervised machine learning [2]. The semi-supervised (weakly supervised) is implemented with little supervision. The idea of this type of machine learning is to use a set of seeds to provide a system with a little external support to start learning how to extract. For example, finding names of the diseases to extract can be done by providing the system with seeds, such as five disease names. First, the system seeks out the sentences that contain these seeds in order to understand the contexts in which they appear. Then, the system tries to find other disease names that exist in the same context [11]. Nadeau [2], developed a semi-supervised NER technique that learned to recognize 100 entity types with little supervision. With regards to unsupervised machine learning, clustering is considered the most typical approach where, for example, named entities can be gathered based on the similarity of context from clustered groups [2], [11].

2.3 Hybrid approach

This approach combines the hand crafted recognition and machine learning methods.

3. Arabic Named Entity Recognition Systems Using Rule-Based

Named Entity Recognition has had an influential role in developing various types of Natural Language Processing (NLP) systems, such as text clustering, Information Retrieval (IR), Question Answering (QA), Machine Translation (MT) and text summarisation, and has served to improve their performance [12]–[15]. Although most research has been devoted to the English language, in the past few years, researchers have started paying attention to the Arabic language. Arabic textual data in electronic form has rapidly increased with more than 20,000 Arabic websites on the Internet and more than 300 million users [16]. Hence, there is a need for tools to deal with this type data, i.e. extracting useful information from Arabic text.

Most of the systems developed for Arabic in the literature focus on NER [12], [17]. The majority of systems developed for Arabic in this field rely on predefined proper name gazetteers. Maloney and Niv [13] developed a system called TAGARAB in order to recognize information relating to names, dates, times, and numerics within Arabic text. A combination of a pattern matching engine and morphological analysis as well as a words list is used to achieve the recognition task. The morphological analysis is used to assist the system in recognizing the various morphological word-shapes and to provide part of speech information for each token before entering the data into the pattern matching engine, i.e. a high precision morphological analysis is employed. The performance achieved for the aforementioned entities is presented in Table 1. Good results were achieved but it is an expensive process as each token is examined. Also, Mesfar [16] developed a system for the Arabic language to recognise proper names (person, location, organization names), dates and numerics through a combination of morphological analysis, syntactic grammar and rules. The role of the morphological analyser is to strip off affixes from inflected words to assist in the matching process. Also, this system relies on gazetteers that contain the names of persons, locations and organizations, together with trigger words that indicate entities of interest, such as person’s title. As in the TAGARAB system, this is an expensive process because the whole text must enter the morphological analyser. Furthermore, huge knowledge resources are used in this system, which include many predefined gazetteers, such as a personal names list that contains 12,400 Arabic first names and a locations list that consists of 5,038 entries, as well as a list of keywords that has 872 trigger words for indicating to these entities. The evaluation results of this system are listed in Table 1. Likewise, Shaalan and Raza [18] developed a system called Name Entity Recognition for Arabic (NERA) to extract 10 named entities from Arabic text; person name, location, company, date, time, price, measurement, phone number, ISBN and file name. They use a rule-based approach that relies on various fixed predefined dictionaries, such as for personal names (263,598 complete names, 175,502 first names and last names with 33,517 names), locations (4,900 names) and organizations (273,491 names of companies). Also, there is a dictionary containing trigger words (indicator words) for helping to identify entities, such as using job titles to indicate persons' names. Moreover, a dictionary called Blacklist is used to reject unwanted entities in order to filter the result. It is noticeable that gazetteers are extensively used but building them is time consuming because Arabic resources (corpora, gazetteers) are generally not free and can be hard to access; they are also relatively few in number. Table 1 presents the performance results achieved for this system in terms of precision, recall, and f-measure.

Moreover, Al-Shalabi et al. [19] presented an algorithm for extracting proper nouns from Arabic texts. They use a set of keywords and special verbs together with some specific rules. Firstly, they use predefined keywords to mark phrases that may contain proper nouns. Secondly, the proposed rules are applied to extract the proper nouns that directly follow the keywords and then the extracted words are classified into one of these categories, based on the type of the keyword: people, locations, organizations, events and products. However, it should be noted that the system is not able to extract proper nouns that do not appear directly after the keywords or the special verbs. Although they reported that they could extract 86.1% of proper nouns in a text, they evaluated their developed system using only 20 documents, which is a very small set of data and is perhaps insufficient for determining the effectiveness of the system. Nevertheless, the performance evaluation is presented in Table 1.

Elsebai et al. [20] adopted a rules-based approach that makes use of the outputs generated by the Buckwalter Arabic Morphological Analyser (BAMA) for developing the Persons Names Arabic Extraction System (PNAES). Their system uses a set of keywords (introductory verbs and words) to indicate the phrases that might contain person names, i.e. there is no predefined person names gazetteer. However, lists containing Arabic person names that start with the definite article "al / the", organisations and location names are used. The reason for using organization and location names lists is to match the extracted word with words in the lists, and once the matching occurs, the word is discarded, otherwise it is classified as a person name, i.e. it is similar to the Blacklist used in [18] to perform filtering. Although this system is able to deal with names that appear not necessarily next to a keyword, unlike the above system developed by Al-Shalabi et al. [19], quite complicated rules are created to cover all the probabilities of a person's name occurring in a text. Finally, the performance achieved for the persons names recognition task was good but it was tested only on one resource, and therefore, the efficiency of this system was not rigorously examined. Table 1 presents the performance results.

Abulei and Evens [21] developed an events extraction and classification system for Arabic information retrieval systems. Their developed system is based on a predefined keyword lists and a parser to extract the events, the dates and related proper nouns. The lists include event words (such as assassinated), valid dates and proper nouns (people, organisations, locations). The system relies on 'direct look-up' to recognise the type of an event, i.e. if an event type is not in the list, it will not be identified.

Additionally, Abueil [22] proposed an algorithm for scanning and understanding events (natural disasters, bombings and deaths) in Arabic text in order to extract information related to them, such as event locations, event types and dates. The system uses different lists that are described as event elements. For example, in order for an event to be marked in the text, special words (keywords), including nouns (e.g. earthquakes, hurricanes and massacres) and verbs, are used. A proper name list, consisting of person, location and organization names, is used to extract entities from the text. Moreover, some particles and nouns are used as link tools when they occur in the text next to proper nouns and noun phrases, linking the event and its location. Also, some particles and nouns are used as relationship tools between more than one event in the text when they occur before the keyword of the second event. The evaluation results of the system were good, as in Table 1, however, the system was only tested on a corpus compiled from one source. Furthermore, the above lists were manually built, based on reading the texts, which is a key disadvantage of this system. In order to confirm this drawback, the system was tested twice; the first experiment showed that some keywords were missing, and some particles and nouns had to be added to the lists. It was reported that after updating the lists, in the second experiment, the system was not able to extract 28 events due to 9 missing keywords. This problem is caused by a lack of any deep contextual analysis, which must be done on huge Arabic texts (within the data collection phase). Also, the authors mentioned that once all the elements (keywords, noun phrases, special nouns and particles, and proper nouns) were evident appear in the description of the event, the system worked well, but this means that the system fails to detect and understand the event if one of element missing.

Piskorski et al. [23] developed a multilingual news event extraction system at the Joint Research Centre of the European Commission. The system was able to extract violent and natural disaster events from online news. They use pattern matching engine and a set of lexicons, which means that a rulebased approach is adopted. The event extraction grammar was originally designed to be applied on the English language but the technique has been extended to work on other languages, such as French, Italian and Arabic. With regards to the Arabic language, the Arabic news articles were translated into English using translation systems, and then they implemented the event extraction grammar. The evaluation results were not reported for this system.

System	Entity	Precision	Recall	F- measure	Year
TAGARAB	Number	82.8	97.0	97.3	1998
	Time	91.0	80.7	85.5	
	Location	94.5	85.3	89.7	
	Person	86.2	76.2	80.9	
Mesfar	Number	97.0	94.0	95,5	2007
	Time	97.0	95.0	96.0	
	Location	82.0	71.0	76.0	
	Person	92.0	79.0	85.0	
Abueil	Event	86	81	84	2007
NERA	Time	97.25	94.5	95.4	2008
	Location	77.4	96.8	85.9	
	Person	86.3	89.2	87.7	
Al-Shalabi	Time	89.4	×	×	2009
	Location	91.6	×	×	
	Person	81.1	×	×	
PNAES	Person	93	86	89	2009
Traboulsi	Person	×	×	×	2009

Table 1. Precision, Recall and F-Measure of Above Systems

Clearly, all the above systems use the rule-based approach. Also, common entities (person names, locations, organizations, dates and numbers) have been investigated by these systems except the two systems developed by Abulei and Evens [21] and by Abueil [22]; they tried to recognise events in texts. Additionally, the described systems did not mention the data sparseness problem in the Arabic language, except [18]. Furthermore, it seems that most dictionaries (gazetteers), especially the keywords lists in the aforementioned systems, were built based on authors' observations or knowledge. In other words, there is no objective explanation or analysis phase carried out on the data being studied for identifying the keywords. However, this phase (data analysis phase) was discussed by Traboulsi [24] in order to identify patterns of person names in Arabic texts. Three types of analysis (frequency, collocation and concordance) were conducted on huge corpora to identify the most important keywords, to discover the most frequent words collocating with keywords, and to obtain the concordance of the keywords. Consequently, the most frequent named entity structures were discovered, which led to the construction of a local grammar for recognising person names, i.e. the other structures that may contain person names are discarded or neglected. As a result, the system might not be able identify of some named entities. Also, no performance evaluation was conducted for this system.

4. Arabic Named Entity Recognition Systems using Machine Learning

On the other hand, machine learning approaches have also been adopted in Arabic NER research. Benajiba et al. [25] designed the Arabic Named Entity Recognition System (ANERsys) based on Maximum Entropy (ME). Annotated corpora were used in this work as well as external resources such as dictionaries. Three different gazetteers were manually built: a location gazetteer consisting of 1,950 names of continents, countries, cities, rivers and mountains, a person gazetteer containing 2,309 names and an organization gazetteer consisting of a list of 262 names of companies, football teams and other organizations. As mentioned earlier, it is time consuming to build gazetteers. Also, in order for the system to be tested, several experiments must be first performed to train it and to derive a set of features for assisting in the recognition process. Moreover, Benajiba and Rosso [26], changed the probabilistic model by using Conditional Random Fields (CRF) instated of ME. Although, they achieved promising results through their improvement of ANERsys, their system still relies on an annotated corpus and the same predefined manually built gazetteers that were used with ME. Preprocessing, such as stemming as well as a part of speech feature is used. Also, they use a nationality feature for marking nationalities in the input text because nationalities are utilised in detecting the named entities; they are used as precursors to recognizing them. This feature relies on a dictionary of 334 different nationalities, which was manually built. Table 2 shows the evaluation results for ANERsys using ME and CRF. Also, Benajiba et al. [14] compared three machine learning approaches: Support Vector Machines (SVM), ME and CRF. The latter was the best and it yielded an overall f-measure of 83.

AbdelRahman et al. [27] integrated two machine learning techniques (bootstrapping semi-supervised pattern recognition and Conditional Random Fields (CRF)) for identifying 10 named entities: person, location, organization, job, device, car, cell phone, currency , date, and time classes. However, thedeveloped system, as with the above systems, relies on prede- fined gazetteers (person (3,228), location (2,183), organization (403), job (70), device (253), car (223) and cell phone (184) to assist in recognising the entities. Also, 16 different features are employed for implementing CRF as well as 232 different seeds. Table II lists the performance results of this system. Abdul-Hamid and Darwish [17], created a system that is able to recognise named entities (person, location and organisation names) for the Arabic language, based on a set of features without using morphological or syntactic analysis or gazetteers. For implementing this work, Conditional Random Fields (CRF) was used. This technique was trained on a large set of surface features in order to avoid using Arabic morphological and syntactic features. Table II presents the evaluation results. It is noticeable that the above machine learning approaches need an annotated corpus for them to be implemented. According to Ku [28], they also need large training data sets. Moreover, they often rely on different predefined gazetteers.

System	Entity	Precision	Recall	F- measure	Year
ANERsys (using ME) (with gazetteers)	Location	82.17	78.42	80.25	2007
	Person	54.21	41.01	46.69	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
(without gazetteers)	Location	82.41	76.90	79.56	
	Person	52.76	38.44	44.47	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
ANERsys (using CRF)	Location	93.03	86.67	89.74	2008
	Person	80.41	67.42	73.35	
	Misc.	71.0	54.20	61.47	
	Organisation	84.23	53.94	65.76	
AbdelRahman (with pattern feature)	Location	96.05	80.86	87.80	2010
	Person	89.20	54.68	67.80	
	Organisation	84.95	60.02	70.34	
(without pattern feature)	Location	89.37	69.25	78.03	
	Person	87.01	53.23	66.05	
	Organisation	88.45	49.00	63.07	
Abdul-hamid and Darwish	Location	93	83	88	2010
	Person	90	75	81	
	Organisation	84	64	73	

Table 2. Precision, Recall and F-Measure for Machine Learning Systems

Table 3 presents a comparison of all the above systems in terms of method, type of corpus and whether or not they use gazetteers, POS and/or stemming.

System	Method	Stemming	gazetteers	POS	Annotated corpus
TAGARAB	Rule based	✓	✓	✓	✗
Mesfar	Rule based	✓	✓	✓	✗
NERA	Rule based	✓	✓	✗	✗
Al-Shalabi	Rule based	✓	✗	✗	✗
PNAES	Rule based	✓	✓	✓	✗
Abueil	Rule based	✓ for keywords	✓	✗	✗
Traboulsi	Rule based	✓	✗	✗	✗
ANERsys	ME	✓only prefixes	✓ ✗	✗	✓
ANERsys	CRF	✓	✓	✓	✓
AbdelRahman	Bootstrapping + CRF	✓	✓	✓	✓
Abdul-hamid	CRF	✓	✗	✗	✓

Table 3. A Comparison Between the System Applied to Applied to Arabic Text

5. Conclusion

In this paper, we have presented named entity recognition systems that were developed for the Arabic language. These systems are divided into two groups based on the approach (rule-based or machine learning) used. The goals, components, performance, strength and weakness of each system are discussed. Also, comparisons are conducted between the above mentioned systems.

References

[1] Fan, W., Wallace, L., Rich, S., Zhang, Z. (2006). Tapping into the power of text mining, *In:* Communications of ACM, 49 (9) 76–82.

[2] Nadeau, D. (2007). Semi-supervised named entity recognition: Learning to recognize 100 entity types with little supervision, Ph.D. dissertation, University of Ottawa.

[3] Michailidis, I., Diamantaras, K., Vasileiadis, S., Frre, Y. (2006). Greek named entity recognition using support vector machines, maximum entropy and onetime, *In:* Proceedings of the 5th International Conference on Language Resources and Evaluation, p. 47–52.

[4] Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists, *In:* Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, ser. EACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, p. 15–21. [Online]. Available: http://portal.acm.org/citation.cfm?id=1609039.1609041

[5] Nezda, L., Hickl, A., Lehmann, J., Fayyaz, S. (2006). What in the world is a shahab? wide coverage named entity recognition in arabic, *In:* Proceedings of the 2006 Language Resources and Evaluation Conference (LREC 2006). Genoa, Italy.

[6] Gaizauskas, R. An information extraction perspective on text mining: Tasks, technologies and prototype applications. [Online]. Available: http://www.itri.brighton.ac.uk/projects/euromap/ Text%20Mining%20Event/Rob Gaizauskas.pdf

[7] Toral, A. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using wikipedia, in In EACL 2006.

[8] Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks, *In:* Proceedings of the eleventh national conference on Artificial intelligence, ser. AAAI'93. AAAI Press, p. 811– 816. [Online]. Available: http://portal.acm.org/citation.cfm?id=1867270. 1867391

[9] Cowie, J., Lehnert, W. (1996). Information extractinon, *Communication of the ACM*, 39 (1) 81–91. [Online]. Available: http://doi.acm.org/10.1145/234173.234209

[10] Nedellec, C. (2002). Machine learning applied to information extraction in specific domains - an example, gene interaction extraction from bibliography in genomics.

[11] Sekine, S. Named entity: History and future. New York University, Department of Computer Science. [Online]. Available: http://cs.nyu.edu//'sekine/papers/NEsurvey200402.pdf

[12] Attia, M., Toral, A., Tounsi, L., Monachini, M., van Genabith, J. (2010). An automatically built named entity lexicon for arabic, *In:* LREC 2010. Valletta, Malta, p. 3614–3621.

[13] Maloney, J. and Niv, M. (1998). Tagarab: A fast, accurate arabic name recognizer using high-precision morphological analysis, *In:* University of Montreal, p. 8–15.

[14] Benajiba, Y., Diab, M., Rosso, P. (2009). Arabic named entity recognition: A feature-driven study, *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, 17 (5) 926–934, July.

[15] Abuleil, S. (2004). Extracting names from arabic text for question-answering systems, *In:* RIAO, p. 638–647.

[16] Mesfar, S. (2007). Named entity recognition for arabic using syntactic grammars, *In:* Natural Language Processing and Information Systems, ser. Lecture Notes in Computer Science, Kedad, Z., Lammari, N., Mtais, E., Meziane, F. and Rezgui, Y., Eds. Springer Berlin Heidelberg, 4592, p. 305–316.

[17] Abdul-Hamid, A., Darwish, K. (2010). Simplified feature set for arabic named entity recognition, *In:* Proceedings of the 2010 Named Entities Workshop, ACL 2010, 2010, p. 110–115.

[18] Shaalan, K., Raza, H. (2008). Arabic named entity recognition from diverse text types, *In:* Proceedings of the 6th international conference on Advances in Natural Language Processing, ser. GoTAL '08. Berlin, Heidelberg: Springer-Verlag, p. 440–451. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-85287-2 42

[19] Al-Shalabi, R., Kanaan, G., Al-Sarayreh, B., Khanfer, K., Al-Ghonmein, A., Talhouni, H., Al-Azazmeh, S. (2009). Proper noun extracting algorithm for arabic language, *In:* International Conference on IT, Thailand.

[20] Elsebai, A., Meziane, F., Belkredim, F. Z. (2009). A rule based persons names arabic extraction system, *In:* Proceedings of the IBIMA,Cairo, Egypt, 11, p. 53–59.

[21] Abuleil, S., Evens, M. (2004). Events extraction and classification for arabic information retrieval systems, *In:* Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence, ser. ICTAI '04. Washington, DC, USA: IEEE Computer Society, p. 769–770. [Online]. Available: http://dx.doi.org/10.1109/ICTAI.2004.59

[22] Abuleil, S. (2007). Using nlp techniques for tagging events in arabic text, *In:* Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence - Volume 02, ser. ICTAI '07. Washington, DC, USA: IEEE Computer Society, p. 440–443. [Online]. Available: http://dx.doi.org/10.1109/ICTAI.2007.186

[23] Jakub, P., Hristo, T., Martin, A., Erik, V. D. G. (2008). Cluster-centric approach to news event extraction, *In:* Frontiers in Artificial Intelligence and Applications, vol. 18. IOS Press, 2008, p. 276–290. [Online]. Available: http://www.iospress.nl/loadtop/load.php?isbn=9781586039042;http://www.booksonline.iospress.nl/Content/View.aspx?piid=10335http://publications.jrc.ec.europa.eu/repository/handle/111111111/4909

[24] Traboulsi, H. (2009). Arabic named entity extraction: A local grammar-based approach, *In:* Proceedings of the International Multiconference on Computer Science and Information Technology, p. 139–143.

[25] Benajiba, Y., Rosso, P., Ruiz, J. (2007). Anersys: An arabic named entity recognition system based on maximum entropy, *In:* CICLing, p. 143–153.

[26] Benajiba, Y. , Rosso, P. (2008). Arabic named entity recognition using conditional random fields, *In:* Proceedings of 2008 Arabic Language and Local Languages Processing Workshop, LREC'08, Marrakech, Morocco.

[27] AbdelRahman, S., Elarnaoty, M., Magdy, M., Fahmy, A. (2010). Integrated machine learning techniques for arabic named entity recognition, *IJCSI International Journal of Computer Science Issues*, 7 (3) 27– 36, July.

[28] Ku, C. H., Iriberri, A., Leroy, G. (2008). Natural language processing and e-government: crime information extraction from heterogeneous data sources, *In:* Proceedings of the 2008 international conference on Digital government research, ser. dg.o '08. Digital Government Society of North America p.162–170. [Online]. Available: http://portal.acm.org/citation.cfm?id=1367832.1367862