A Novel Approach for Detecting Arabic Persons' Names using Limited Resources

Omnia Zayed, Samhaa El-Beltagy, and Osama Haggag

Center of Informatics Science, Nile University, Giza, Egypt

{omnia.zayed, samhaaelbeltagy, osama.haggag}@gmail.com

Abstract. Named entity recognition is an involved task and is one that usually requires the usage of numerous resources. Recognizing Arabic entities is an even more difficult task due to the inherent ambiguity of the Arabic language. Previous approaches that have tackled the problem of Arabic named entity recognition have used Arabic parsers and taggers combined with a huge set of gazetteers and sometimes large training sets. However, the recent surge in the usage of social media, where colloquial Arabic, rather than modern standard Arabic is used, invalidates these approaches because existing parsers fail to parse colloquial Arabic at an acceptable level of precision. To address such limitations, this paper presents an approach for recognizing Arabic persons' names without utilizing any Arabic parsers or taggers. The approach uses only a limited set of publicly available dictionaries. The followed approach integrates dictionaries with a statistical model based on association rules for extracting patterns that indicate the occurrence of persons' names. Through experimentation on a benchmark dataset, we show that the performance of the presented technique is comparable to the state of the art machine learning approach.

Keywords: Arabic Named Entity Recognition, Association Rules, Colloquial Arabic, Modern Standard Arabic.

1 Introduction

The importance of named entity recognition (NER) is increasing progressively due to the necessity of a better understanding of human communication. A lot of applications in the field of Natural Language processing make use of NER as extensively listed in [12]; examples of those applications include Machine Translation, Text Clustering and Summarization, Information Retrieval and Question Answering systems.

Approaches for recognizing named entities from text, fall under three categories. The first approach known as "rule based NER" combines grammar, in the form of handcrafted rules, with gazetteers to extract named entities. The second, is "machine learning based NER" which utilizes large datasets and features extracted from these, to train a classifier to recognize a named entity. Hence this approach converts the named recognition task into a classification task. Machine learning algorithms could be further categorized as either supervised or unsupervised. The third and final ap-

proach is "hybrid NER" which combines both of the aforementioned approaches [18, 25]. A comparison between rule based approaches and machine learning approaches is presented in [18] in terms of the used domain. The reason behind the difficulty of modifying rule based approach for new domains has been related to the use of a lot of resources such as gazetteers, besides the need of complicated linguistic analysis to detect the named entities. On the other hand, machine learning approaches need a training dataset which is tagged in a certain manner to recognize new entities from new testing dataset of the same domain. Besides a precise selection of features is required [1, 18, 25].

Building a system to extract Arabic named entities is a difficult task. Being a Semitic language, the Arabic language is well known for its complex morphology. In addition, Arabic has a unified orthographic case; it does not have capital letters. Conversely, in the English language which allows mixed letter cases; some named entities can be distinguished because they are capitalized. These include persons' names, locations and organization. Moreover, Arabic is notable for its inherent ambiguity in which one word could imply variety of meanings [17, 25]. The fact that many names are derived from adjectives complicates the task of recognizing persons' names even further. The distinctive challenges of Arabic language including ambiguity and complexity are explained in detail in [1].

While most existing Arabic texts are written in formal Modern Standard Arabic Text (MSA), the volume of informal colloquial Arabic text is increasing progressively with the wide spread use of social media examples of which are Facebook, Google Moderator and Twitter. Previous approaches that have tackled the problem of Arabic named entity recognition frequently depend on Arabic parsers and taggers combined with a huge set of gazetteers and sometimes large training sets to achieve their task. However, the task of named entities extraction from colloquial Arabic text invalidates these approaches as existing parsers fail to parse colloquial Arabic at an acceptable level of precision. This is due to sentence irregularity, incompleteness and the varied word order of colloquial Arabic. Colloquial Arabic also has no standard rules or grammatical constructions because it maps to a spoken language [25].

To address such challenges, this paper introduces an approach to recognize Arabic persons' names without utilizing any Arabic parsers or taggers. Moreover the followed approach tries to overcome the ambiguity problem of persons' names by organizing publicly available dictionaries of person names into clusters as will be detailed. Since the presented approach makes use of a limited set of dictionaries, integrated with a statistical model based on association rules, the model can easily generalize to different domains in our future work.

The rest of this paper is organized as follows: section 2 describes the proposed approach in detail. In section 3, system evaluation on a benchmark dataset is discussed. Section 4 reviews an overview of the literature on NER systems in Arabic language. Finally conclusion is presented in section 5.

2 The Proposed Approach

In this work, a rule based approach combined with a statistical model, is adopted in a novel way to identify and extract person names from Arabic text. Our approach tries to overcome two of the major shortcomings of using rule based techniques which are the difficulty of modifying a rule based approach for new domains and the necessity of using huge set of gazetteers. The proposed approach builds a statistical model for automatically extracting patterns which indicates persons' names occurrences, scored using association rules. Moreover, the ambiguity problem of persons' names is overcome using a clustering technique. Our approach consists of two phases, as shown in Figure 1. In the first phase, "The building of resources phase", person names are collected and clustered, and "name indicating" patterns are extracted. In the second phase, "Extraction of persons' names phase", name patterns and clusters are used to extract persons' names from input text. Both of these phases are described in depth, in the following subsections.

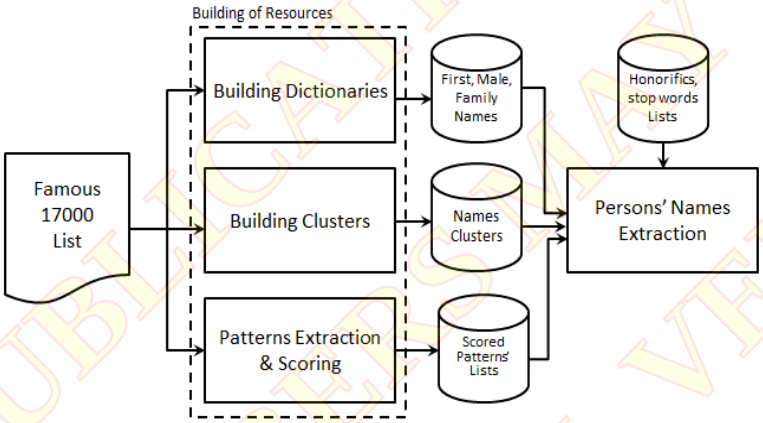

Fig.1. System Architecture.

2.1 The Building of Resources Phase

In this phase the resources on which the system depends are prepared. This phase is divided into 4stages.In the first stage, persons' names are collected from public resources. In the second stage, dictionaries of first, middle/male and family persons' names are built from collected resources. In the third stage, names are grouped together into clusters to avoid the Arabic persons' names ambiguity problem as will be detailed later. In the fourth and final stage, a corpus is used to build and score patterns which indicate the occurrence of a person name. Scoring of the patterns is done using association rules.

Persons' Names Collection. Wikipedia, with its huge collection of names under the people category, offers an excellent resource for building a name database. Kooora, which is an Arabic website for sports, also provides a large list of names. So,

in this stage, Wikipedia[1] and Kooora[2] websites were used to collect a list of about 17,000 persons' names. Since the aim of this work is not just to recognize names of famous people, but instead to identify the name of any person even if it does not appear in the collected list, the collection is further processed and refined in order to achieve this goal in the following step.

Building of Dictionaries. In this stage, the list of names collected in the previous step (we call this list the "famous_17000_list".) is processed in such a way so as to separate first names from family names. The processing step includes handling the different variations of Arabic persons' names. As described in [25], Arabic name could have affixes such as prefixes or embedded nouns. A word preceded or followed by those affixes must not be split on white spaces, instead the word and its affix should be considered as a single entity. For example, the male name عبدالعزيز (abdulaziz) should not be split as عبد (abd) as first name and العزيز (alaziz) as family name, instead it should be treated as single entity عبدالعزيز (abdulaziz) and considered as a first name. Table 1 lists the different variations of Arabic persons' names with examples [25].

Table 1. Different variations of writing Arabic persons' names.

Case	Example
Simple case (no affixes)	احمد محمود Ahmad Mahmoud
Prefix case {عبدAbd, ابو Abou, بن Bin, ال Al, …etc}	**عبد** العزيز **ال** سعود Abdulaziz Al Saud
Double prefix case {ابوعبدAbou Abd,بن عبد Bin Abd, … etc}	سلطان**بن عبد** العزيز **ال** سعود Sultan bin Abdulaziz Al Saud
Embedded noun case {الدين El-Deen,الله Allah, …etc}	هيردي نور **الدين** Herdi Noor Al-Din
Complex name (prefix + embedded noun)	تقي **الدين** محمد **بن** معروف الشامي Taqi al-Din Muhammad Ibn Ma'ruf al-Shami

Building of Name Clusters. Once names dictionaries are built, they can be used to identify previously unseen names by stating that a full name is composed of a first name followed by other male names and/or a family name. However, the problem is not that simplistic. One of the problems of rule based NER systems is that straight forward matching of persons' names using dictionaries, can often result in mistakes. For example, a phrase such as في خطاب بوش (In Bush's speech) a full name could be mistakenly extracted as خطاب بوش (Khatab Bush) even though it is highly unlikely that an Arabic person's name such as خطاب (Khatab) will appear besides an American person's name such as بوش (Bush). Arabic text often contains not only Arabic names, but names from almost any country transliterated to Arabic.

[1] http://ar.wikipedia.org/wiki/تصنيف:تراجم

[2] http://www.kooora.com/default.aspx?showplayers=true

The "famous_17000_list" thus contains Arabic, English, French, Hindi, and Asian persons' names, written in Arabic language. In our approach, clustering is used to separate these names. Clustering this list is an important step to determine acceptable name combinations. To carry out clustering, we have used the Louvain [13] graph clustering technique from within Gephi [6] which is an open source software for exploring and manipulating networks.

As a pre-processing step, the 17,000 persons' names list processed to build a dictionary in which each first name is a key item whose corresponding value is a list of the other family names it had, accounting for redundancies. The variations of writing Arabic persons' names mentioned in the previous sub-section is considered. This dictionary is converted to a graph, such that each first name and family name form separate nodes. Edges are then established between each first name and its corresponding family names. The resulting graph consisted of 15782 nodes, and 16481 undirected edges.

Then, the Louvain method was applied to the graph for finding communities within the network. The community in this context is a cluster of names that are related. A resolution parameter of 3.5waschosen, allowing larger communities to be found. The outcome was a set 2116 clusters, where each name is given a modularity class number denoting which community (cluster) it belongs to.

Figure 2 shows a snapshot of the resulting clusters. It was observed from visualizing the data that most of the culturally similar names were grouped together, for example it can be noted that most of the names common in the Arabic-speaking regions were grouped together. The same applies to English and French names and to other names that are kind of unique to their region. It was also observed that small lone clusters are those that contain rare names that do not have connections to the other names.

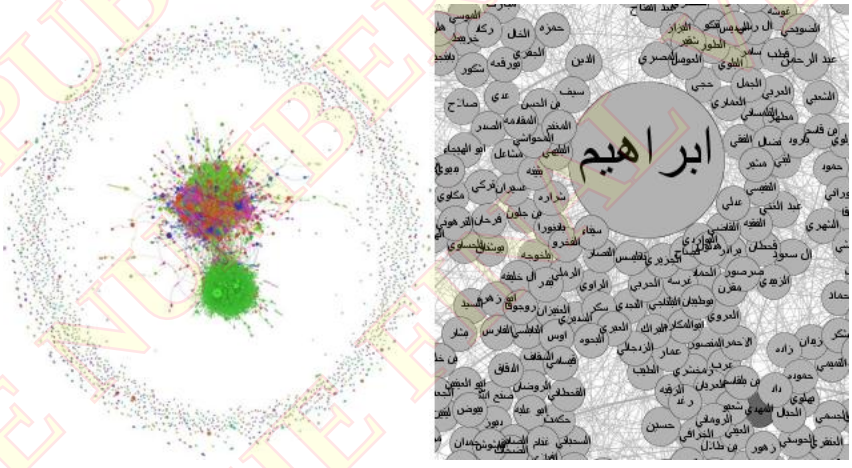

Fig.2. Visualization of generated clusters, to the left are all generated, lone clusters can be seen on the border and the two largest clusters are those of Arabic names (below) and Western names (above). To the right is a closer view of a subset of the Arabic names cluster.

Extracting Scored Patterns. In this stage, a statistical model is built to automatically learn patterns which indicate the occurrence of a person name. This stage is divided into 4 steps, as shown in Figure 3.

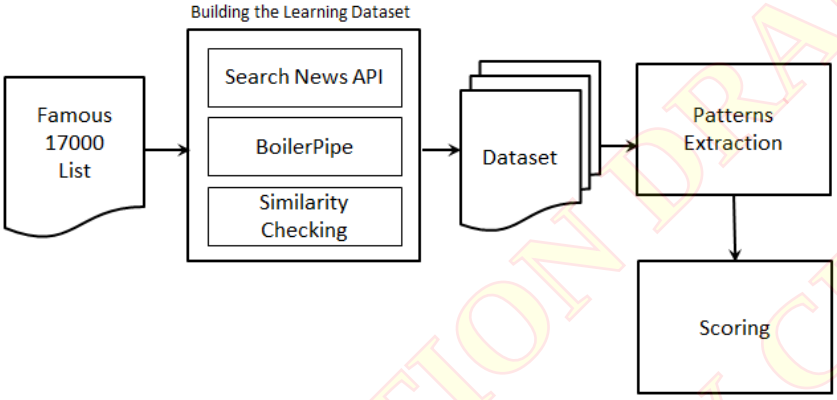

Fig.3. Building lists of scored patterns stage

Initially each name in the"famous_17000_list" is used as a query to search news articles to build learning dataset from the same domain that we are targeting to extract persons' names from. Akhbarak[3] API and Google Custom Search API[4]were used to search and retrieve news stories.

Around 200 news article links are crawled for each person name in the "famous_17000_list. After downloading the pages associated with these links BoilerPipe[5] is used to extract the main news article. Often news stories are repeated in many sources, so very similar stories are detected and removed.

Following this step, unigram patterns around each name are extracted. Three lists are formed. A complete pattern list keeps set of complete patterns around the name with their count. A complete pattern consists of <$word_1$><name><$word_2$>. The <name> part just indicates that a name has occurred between words: $word_1$ and $word_2$. Two type of unigram pattern lists are kept: a "before" list keeps the patterns that appear before a name with their counts (example: اكد (confirmed)) and an "after" list stores patterns that occur after a name with their count (example:ان (that)).

Finally the support measure employed by association rules [5] is used to score each pattern in the three lists. Support is calculated as the ratio of the count of a pattern followed by a name over the total count of all patterns followed by a name. For example the support rule used to score a unigram pattern before a name is computed by the following equation.

$$support_{patternbeforename} = \frac{Countthispatternfollowedbyaname}{Countoftotalsumofallpatternsfollowedbyaname} \quad (1)$$

[3] http://www.akhbarak.net/
[4] https://developers.google.com/custom-search/v1/overview
[5] http://code.google.com/p/boilerpipe/

The newly created three lists of scored patterns are saved descendingly according to the value of the score.

2.2 Extraction of Persons' Names Phase

The persons' names extraction process is dependent on the previous pre-prepared resources which are the dictionaries of first, and family names, the name clusters, a list of honorifics, a list of stop words and the patterns lists. Rules are implemented to extract persons' names from the unseen dataset of the same targeted domain. The benchmark dataset, ANERcorp[6]is used to evaluate the proposed system. The system assumes that any full name consists of a first name followed by one or more male names followed by zero or one family name. The generated name clusters are used to ensure that all candidate portions of a name fall in the same cluster to avoid matching mistakes. One of the rules used in the extraction phase is as follows:

```
For each word wi in the target text:
  If wi in patterns_before_list
    If wi+1 in honorific_list
    Check for names from wi+2 in the same cluster;
  Stop when a delimiter d is_found where d ∈
  (pattern_after|stop_word|punctuation|title_start)
    Else
  Check for names from wi+1in the same cluster;
  Stop when a delimiter d is_found where d ∈
  (pattern_after|stop_word|punctuation|title_start)
  Else if wi in honorific_list
  Check for names from wi+1 in the same cluster;
    Stop when a delimiter d is_found where d ∈
(pattern_after|stop_word|punctuation|title_start)
```

The above rule is used to extract names from sentences such as:

قال *الرئيس* محمد مرسي **ان** مصر تخطو ...

President Mohammad Morsi said that Egypt is stepping through …

اكد *الدكتور* محمد حجازي *مساعد* وزير الخارجية ...

Dr. Mohammad Higazy, Deputy of minister of foreign affairs confirmed that …

قال وليد جنبلاط *رئيس* كتلة اللقاء ...

Walid Junblatt the president of … said…

This rule is generalized to extract names from sentences which contain multi honorifics before the person's name such as:

قال *رئيس الوزراء الاسرائيلي* ايهود اولمرت **إنه** عازم ...

Prime Minister of Israel Ehud Olmert said that he will …

Another rule is used to check for a pattern followed by an unknown name (not in the dictionaries) with the prefix عبد (Abd) followed by known male nameand/or fami-

[6] http://www1.ccls.columbia.edu/~ybenajiba/downloads.html

ly name (the previous stopping criterion is used). Also we utilize the fact that some names could appear with the conjunction و (and), hence a rule is used to extract a known first name preceded by و (and) followed by known male name and/or family name(the previous stopping criterion is used).Other rules are employed, but are not included due to space limitations.

3 System Evaluation

The presented system was evaluated using the precision, recall and f-score measures based on what it extracted as names from the benchmark ANERcorp dataset. Table 2 provides a comparison between the results of the presented system with two state of the art systems which are the hybrid NERA approach [1] and the machine learning approach using conditional random fields (CRF) [10].

Table 2. Comparison between our system performance in terms of precision, recall and F-score with the current two state of the art systems

	Precision	Recall	F-score
Hybrid System	94.9	90.78	92.8
CRF System	80.41	67.42	73.35
Our System	**92.29**	**72.75**	**81.36**

From this comparison, it can be inferred that our system competes with the state of the art systems using the introduced novel approach in terms of precision. However the recall of our system is still below the recall of the state of the art hybrid approach. One of the reasons that decrease the recall is the fact that our main rule is to extract full name which is composed of a first name followed by other male names and/or a family name, in order to avoid the effects of false positives. So we did not extract single names such as ميسي (Messi) and رونالدينيو (Ronaldinio) in the phrase:

... لكن نجميه رونالدينيو والارجنتيني ميسي اضاعا...

…but its stars Ronaldinio and the argentine Messi have missed …

Another reason is that, using our dictionaries in addition to the above rule, a name such as ابنيزر نجوي (Ebenezer Najwa) could not be extracted, because according to our dictionaries (Najwa) is a first name so it should not be found after (Ebenezer).

To handle the issue of the relatively low recall, we will try to make the used rules more flexible while keeping an acceptable precision value.

Table 3 shows the effect of using clusters in boosting the system precision besides the final system results using patterns and clusters. The use of clusters decreases the effect of false positives so the precision increased by 6points.

Table 3. Individual system‘s components evaluation vs. the complete system in terms of precision

	Precision	Recall	F-score
Dictionaries Only	71.0	62.98	66.75
Dictionaries+Clusters	77.24	58.62	66.65
Dictionaries+Clusters+Patterns	**92.29**	**72.75**	**81.36**

4 Related Work

NER systems for Arabic can be classified based on the type of text being processed; whether it is formal Modern Standard Arabic (MSA) or informal colloquial Arabic text. Then they can be further classified based on the used approach, whether it is rule based, machine learning based or hybrid.

The majority of previous work addressing NER in Arabic language was developed for formal MSA text which is the literary language used in newspapers and scientific books. However the informal colloquial Arabic, corresponding to the spoken dialectic, is currently being used widely in social media communication.

As mentioned earlier, rule based NER based on handcrafted rules combined with gazetteers. One of the initial systems which combined a generic pattern matching engine with high-precision morphological text analysis to recognize Arabic named entities was TAGARAB [17]. TAGARAB depends on a morphological analysis module plugged into a morphological tokenizer in addition to lists of nouns, verbs, and adjective stems to partially support a series of regular expressions.

Another technique was presented to extract proper names from Arabic text for a question-answering system [3]. The technique depends on collecting information about the words in the text and building graphs to represent the relationships between them.

A system based on local grammar (patterns) to extract persons' names from Arabic news articles is described in [24]. The used approach is based on the fact that persons' names cluster around certain frequent verbs in news articles. Collocation analysis is done to discover the words that frequently collocate with the Arabic reporting verbs such as prepositions, punctuations and function words. Then concordances analysis is generated to return frequency information and citations for the searched reporting verbs and each of its inflected forms. Finally a Finite State Automata (FSA) is constructed for the reporting verbs extracted patterns.

PERA [21] is a system for extracting Arabic persons' names. The system adopts a rule based approach using linguistic grammar-based techniques. Grammar rules, supported by gazetteers, were built based on keywords or trigger words to form a window around a person's name. PERA was evaluated on purpose-built corpora using ACE and Treebank news corpora that were tagged in a semi-automated way. The system has been generalized as NERA [22, 23] to extract other named entities.

The work presented in [15] describes a person named entity recognition system for the Arabic language. The system makes use of heuristics to identify person names and is composed of two main parts: the General Architecture for Text Engineering (GATE) environment and the Buckwalter Arabic Morphological Analyzer (BAMA). The system makes use of a huge set of dictionaries. The same work was repeated in [16] and compared with [21] and [19].

As mentioned in [1], the frequently used approach for NER is the machine learning approach by which text features are used to classify the input text depending on an annotated dataset.

Benajiba et al. applied different machine learning techniques [7-12] to extract named entities from Arabic text. The best performing of these makes use of optimized feature sets [10].

ANERSys [7] was initially developed based on n-grams and a maximum entropy classifier. The maximum entropy classifier basically computes for each word the probability that it will be assigned to each of the considered classes using the maximum entropy formula and then assigns the class with the highest probability to this word. Moreover a training and test corpora (ANERcorp) and gazetteers (ANERgazet) were developed to train, evaluate and boost the implemented technique. ANERcorp is currently considered the benchmark dataset for testing and evaluating NER systems.

ANERSys 2.0 [8] basically improves the initial technique used in ANERSys by combining the maximum entropy with POS tags information. Hence the recognition of long named entities is improved by extracting the boundaries of the named entity.

By changing the probabilistic model from Maximum Entropy to Conditional Random Fields the accuracy of ANERSys is enhanced [9]. Another system is introduced which makes use of leading and trailing character n-grams in words in addition to other surface and word association features to train a conditional random field's model [2].

A novel approach is described in [11] to extract Arabic named entities using Support Vector Machines with the aid of contextual, lexical and morphological features combination.

A recent attempt to extract named entities from Arabic text using an artificial neural network is discussed in [20]. The system uses a back propagation training algorithm in addition to selecting an appropriate set of features for each named entity class.

Hybrid approaches combine machine learning techniques, statistical methods and predefined rules. In [4] a hybrid system built based on both statistical methods and predefined rules to extract Arabic named entities, is described. The system combines three different techniques: rules, graphs, and statistics. Rules are used to mark named entities phrases. A graph-based method is implemented to represent the relationships between words. Finally rules and the frequency of tokens are utilized to identify proper names.

The most recent hybrid NER system for Arabic uses a rule based NER component integrated with a machine learning classifier [1]. The system operates over two stages. In the first stage, a re-implementation of the NERA system [22, 23] using the GATE platform, is used to tag the words of the input text. In the second stage, the outputs of the rule based system are propagated as features to a decision-tree machine learning classifier along with other general features. The Stanford POS Tagger has been used to compute some of these other features, such as word category and affixation. The reported results of the system are significantly better than the pure rule based system and the pure machine learning classifier. In addition the results are also better than the state of the art Arabic NER system based on conditional random fields [10].

4.1 Differences Between our System and Previous Work

From the previous discussion, it can then be inferred that, the currently used rule based approaches to extract named entities from MSA text, are dependent on tokenizers, taggers and parsers combined with a huge set of gazetteers. Although, those approaches might be sufficient for extracting persons' names from a formal domain, it will be hard to modify them for the informal "colloquial Arabic" domain. Similarly, machine learning approaches make use of taggers, parsers and set of gazetteers to extract contextual, lexical and morphological features. Those features are used to train different classifiers. In addition, an annotated corpus is always required for training.

Our approach, which combines a rule based approach with a statistical one, avoids the use of parsers, taggers and morphological analyzers. All the system requires is a large set of names, which can be easily obtained from public resources such as Wikipedia. The main challenges addressed by this work are to overcome the ambiguity problem of persons' names, to avoid the shortcomings of both the rule based NER and the machine learning based NER approaches and to build a domain independent persons' names extraction system.

There is some similarity between our approach and the one based on local grammar [24] as later uses reporting verbs as patterns to indicate the occurrence of persons' names. However our approach extracts patterns automatically from the same domain under study, so the patterns are not limited to a list of reporting verbs.

There was an attempt to recognize named entities from documents written in Indonesian language using association rules [14]. The system uses association rules in terms of support and confidence to extract named entities. A set of previously defined features, dictionaries and name classes from an annotated corpus is employed to describe the two sets of items from the dataset which are used to calculate the support. One of those sets is the name of the class to be predicted and the other is all the possible forms of the class, but the overall approach taken by that system is different than the one we presented.

5 Conclusion

This paper presented a novel approach for extracting persons' names from Arabic text. This approach integrated name dictionaries and name clusters with a statistical model based for extracting patterns that indicate the occurrence of persons' names. The used approach overcomes major limitations of the rule based approach which are the need of huge set of gazetteers and domain dependence. Using this system, persons' name extraction could be applied on new domains without facing difficulties to import the system into the new domain. Our rule based approach was able to overcome the ambiguity of Arabic persons' names using clusters besides the original dictionaries of names. Building the patterns' statistical model using association rules improved the tasks of Arabic persons' names disambiguation and extraction from any domain. System evaluation, on the benchmark dataset, showed that the performance of the presented technique is comparable to the state of the art machine learning ap-

proach. However, it still needs some improvements to compete with the state of the art hybrid approach.

References

1. Abdallah, S., Shaalan, K.,Shoaib, M.: Integrating rule-based system with classification for Arabic named entity recognition. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7181, pp. 311–322. Springer, Heidelberg (2012)
2. Abdul Hamid, A., Darwish, K.: Simplified feature set for Arabic named entity recognition. In: Proceedings of the 2010 Named Entities Workshop, pp. 110–115. Association for Computational Linguistics, Uppsala (2010)
3. Abuleil, S.: Extracting names from Arabic text for question-answering systems. In: Proceedings of Coupling approaches, coupling media and coupling languages for information retrieval, RIAO 2004,pp. 638-647. Avignon(2004)
4. Abuleil, S.: Hybrid system for extracting and classifying Arabic proper names. In: Proceedings of the fifth WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED 2006, pp. 205-210. Madrid (2006)
5. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, SIGMOD 1993, pp. 207–216. Washington (1993)
6. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM 2009, pp.361–362. California (2009)
7. Benajiba, Y., Rosso, P., Benedì Ruiz, J.M.: ANERsys: An Arabic named entity recognition system based on maximum entropy. In: Gelbukh, A. (ed.) CICLing 2007. LNCS, vol. 4394, pp. 143–153. Springer, Heidelberg (2007)
8. Benajiba, Y., Rosso, P.: Anersys 2.0: Conquering the ner task for the Arabic language by combining the maximum entropy with pos-tag information. In: IICAI, pp. 1814–1823 (2007)
9. Benajiba, Y., Rosso, P.: Arabic named entity recognition using conditional random fields. In: Workshop on HLT & NLP within the Arabic World. Arabic Language and Local Languages Processing: Status Updates and Prospects (2008)
10. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition using optimized feature sets. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, pp. 284–293. Association for Computational Linguistics, Morristown (2008)
11. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: An svm-based approach. In: The International Arab Conference on Information Technology, ACIT 2008 (2008)
12. Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: A feature-driven study. IEEE Transactions on Audio, Speech, and Language Processing 17(5), 926–934. (2009)
13. Blondel, V.D., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment p10008, (2008)
14. Budi, I., Bressan, S.: Association rules mining for name entity recognition. In: Proceedings of the Fourth International Conference on Web Information Systems Engineering, WISE 2003, pp. 325–336. Italy (2003)

15. Elsebai, A., Meziane, F., Belkredim, F. Z.: A rule based persons names Arabic extraction system. Communications of the IBIMA 11, 53–59.(2009)
16. Elsebai, A., Meziane, F.: Extracting person names from Arabic newspapers. In: Proceedings of the International Conference on Innovations in Information Technology, IIT 2011, pp.87–89.UAE (2011)
17. Maloney, J., Niv, M.: TAGARAB: a fast, accurate Arabic name recognizer using high-precision morphological analysis. In: Proceedings of the Workshop on Computational Approaches to Semitic Languages, Semitic 1998, pp. 8–15. Association for Computational Linguistics, Morristown (1998)
18. Mansouri, A., Affendey, L.S., Mamat, A.: Named entity recognition using a new fuzzy support vector machine. In: Proceedings of the 2008 International Conference on Computer Science and Information Technology, ICCSIT 2008, pp. 24–28. Singapore (2008)
19. Mesfar, S.: Named entity recognition for Arabic using syntactic grammars. In: NLDB 2007, LNCS, vol. 4592, pp. 305-316. Springer, Heidelberg (2007)
20. Mohammed, N.F., Omar, N.: Arabic named entity recognition using artificial neural network. Journal of Computer Science 8(8), 1285-1293. (2012)
21. Shaalan, K., Raza, H.: Person name entity recognition for Arabic. In: Fifth Workshop on Important Unresolved Matters, pp. 17–24.Czech Republic(2007)
22. Shaalan, K., Raza, H.: Arabic Named Entity Recognition from Diverse Text Types. In: Nordström, B., Ranta, A. (eds.) GoTAL 2008. LNCS (LNAI), vol. 5221, pp. 440–451. Springer, Heidelberg (2008)
23. Shaalan, K., Raza, H.: NERA: Named entity recognition for Arabic. Journal of the American Society for Information Science and Technology, 1652–1663 (2009)
24. Traboulsi, H.: Arabic named entity extraction: A local grammar-based approach. In: Proceedings of the International Multiconference on Computer Science and Information Technology, vol. 4, pp. 139–143 (2009)
25. Zayed, O.H., El-Beltagy, S.R.: Person Name Extraction from Modern Standard Arabic or Colloquial Text. In: Proceedings of the eighth International Conference on Informatics and Systems, INFOS 2012, pp. NLP-44–NLP-48.Egypt (2012)