A Named Entity Recognition System Applied to Arabic Text in the Medical Domain

Saad Alanazi[1 2], Bernadette Sharp[2] and Clare Stanier[2]

[1] College of Computer Science and Information, Aljouf University, Skaka, Saudi Arabia

[2] Faculty of Computing, Engineering and Technology, Staffordshire University, Beaconside, Stafford ST18 0AD, UK

Abstract

At the sixth Message Understanding Conference (MUC-6) in 1995, Named Entity Recognition (NER) was recognised as an essential sub field of information extraction and as an important contribution to natural language processing. The goal of NER is to extract specific predefined list of entities, which can include proper names, numerical expression and temporal expression. This paper introduces NAMERAMA which is a novel NER system based on Bayesian Belief Network (BBN). It extracts disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic text in the medical domain. The results of the developed system shows that BBN performance is promising with 71.05% overall F-measure. The highest F-measure score was achieved in recognising disease names with 98.10% while the lowest was in recognising symptoms with 41.66%.

Keywords: *Named Entity Recognition, Bayesian Belief Network, Natural language processing, Machine learning.*

1. Introduction

The first named entity resolution (NER) was presented by Rau [37] whose system extracted company names computationally from financial news. At the sixth Message Understanding Conference (MUC-6) in 1995, NER was recognised as an essential subfield of natural language processing. The task of NER is to extract specific predefined list of entities such as proper names (e.g. organisations, people, location), numerical expression (e.g. monetary values) and temporal expression (e.g. dates, times). Whilst the study of NER by Petasis et al. is limited to "identifying and semantically tagging proper nouns (PNs) in running texts" [36], other studies extend entity recognition to include the extraction of more complex entities such as chemical substances (e.g. compounds, reagent, solvents) [21], and the identification of disorder named entities from electronic medical records [17]. NER plays an important role not only in assisting machine translation, improving information retrieval process and extracting information [25], but also in populating knowledge bases, supporting web mining and semantic web communities [38] and more recently for law enforcement applications [19].

The aim of this research project is to develop a novel Named Entity Recognition (NER) method to extract cancer disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic texts in the medical domain. This paper starts by reviewing previous work and then describes our proposed approach. A discussion of the problems and issues encountered in implementing our NER approach are also presented.

2. Challenges of Arabic language processing

The literature review shows that most research efforts on NER have been devoted to English language texts, and a good number of studies have been related to German, Spanish and Dutch NER research. Recently, interest has been growing in NER research projects focusing on Arabic texts [32]. Since the Arabic language is the mother tongue of more than 300 million citizens in more than 25 countries, devoting more research to develop NER systems dedicated for the Arabic language is significant [43]. However, Arabic has many traits which make building an effective NER system a very challenging task. Some of these challenges are described below:

2.1 Lack of capitalisation

With languages like English, there is the use of a capital letter and most named entities begin with this capitalisation, making the extraction of proper names a lesser challenge. However, in the Arabic language capitalisation does not exist as an orthographic feature [22]. Furthermore, the lack of capitalisation makes it hard to distinguish most Arabic proper nouns from common nouns and adjectives. Therefore, since ambiguous words are more likely to be used as proper nouns in a text, relying alone on looking up entries in a proper noun dictionary would not be appropriate in tackling these problems [5].

2.2 Agglutination

The Arabic language has an agglutinative nature and this has an outcome of different patterns which can create many lexical variations. It has a very systematic, but complicated morphology. This is seen with words that consist of prefixes, a stem or a root, and sometimes even more than one, as well as suffixes with different combinations. There are also clitics, which in most languages, as well as English, are treated as separate words, but in the Arabic language they are agglutinated to words [23].

2.3 Short vowels absence

Diacritics can be found in the Arabic text which is a representation of most vowels which affect the phonetic representation. This would give an alternative meaning to the same word. Consequently disambiguation in the Arabic language is a difficult task due to the fact that it is written without diacritics [7].

3. Related work

The first NER system used hand-crafted heuristic rules and combined heuristics, exception lists and extensive corpus analysis [37]. In the last two decades, a diversity of algorithms has been applied to NER, with differing strengths and weaknesses. This section reviews the three main NER approaches applied to the Arabic domain: rule based, machine learning and hybrid methods.

3.1 Rule based method

This method depends on hand-made linguistic rules (such as grammar) which are defined by linguists. It has been used extensively in many studies which have adapted the rule based method to extract predefined named entities from social media [52] and from newspapers such as Al Hayat [29], the Aljazera website [20], the Al-Raya newspaper [8], the Assabah and Alanwar newspapers [50] and ANERcorp, which comprises 136 newspaper articles [15, 2, 50, 6, 45]. Other domains of NER application are also studied, namely the financial domain [47], the criminal domain [11], the sport, economic and political domains [4], and the medical domain [40]. The common entities found in most studies include person names, location, organisation, time and date. In some research such as the one conducted by Al-Shalabi et al. [8], NER incorporates diverse entities, such as events and equipment names.

3.2 Machine learning method

Machine Learning (ML) is another approach extensively used to develop statistical models for named entities prediction. The machine learning approaches applied in the literature to Arabic texts are supervised learning approach and semi-supervised approach.

3.2.1 Supervised learning

Supervised Learning (SL) includes studying and analysing both positive and negative features of named entity examples from a broad collection of annotated corpora and also the formation of rules which can capture occurrences of any given type. Techniques that belong to SL are Maximum Entropy Models (ME), Conditional Random Fields (CRF), Support Vector Machines (SVM) and Neural Networks (NN); these have been applied to Arabic texts. Maximum Entropy (ME) has been applied by [15] whose F-measure is 55.23% using person names, location, organisation, and "miscellaneous" whereas [33] has extracted 18 various named entities with an F-measure of over 85%. Conditional Random Fields (CRF) have been applied by [13, 3, 16]. Different corpora were used in order to evaluate their systems, such as ANERcorp and AC 2005. Support Vector Machine (SVM) has been implemented by [14, 27, 34]. Benajiba et al [12] studied the ramifications of using different features with models such as SVM, ME, and CRF, and concluded that both SVMs and CRFs outperformed the ME model. It was also noted that the choice of features is a very significant phase of any ML-based system. Mohammed and Omar [30] adapted neural networks in their approach, making use of the back propagation training algorithm.

3.2.1 Semi-supervised learning

"Bootstrapping" is considered the main technique of semi-supervised learning (SSL) where supervision is minimal and is only included for the start of the learning process. For instance, if the aim is "disease names" the system may ask the user to supply some examples. Then, with the given examples the system will search for sentences and try to recognise possible contextual clues provided by the examples. The system will conduct

another search to find more sentences which have a similar context. The learning process can then be reapplied to the sentences which have been found so that new and relevant contexts can be identified. Through repetitions of this process the system will recognise a broad range of disease names as well as context [31]. Few research efforts have been made to adapt semi-supervised learning methods to Arabic texts. AbdelRahman et al. [2] combined CRF with bootstrapping to extract a wide range of entities which include person, location, organisation, job, device, car, cell phone, currency, date, and time. Their system found that the F-measure varies between 69.47% and 96.05% depending on the type of entities. Althobaiti et al [10] adapted the bootstrapping algorithm in their system (ASemiNER) in order to identify specific entities such as person, location, and organisation scoring 64.14%, 73.06% and 54.52% of F-measure respectively. The system can also recognise specialised entities such as the names of politician, sport personalities and artists.

3.3 Hybrid method

The hybrid method is a combination of the rule based method and the machine learning approach. Over the past few years, some hybrid systems have been established in order to improve the performance of rule based systems and machine learning systems. Abdallah et al. [1] extended the developed the NERA system developed by Shaalan and Raza [44] by combining decision trees with the rule based method. The system recognises three entities which are person, location and organisation. The overall average of the F-measure was 88.87%. Another system developed by Oudah and Shaalan's [35] combined SVMs and Logistic Regression and increased the number of named entities from 3 to 11 types to include person names, location, organisation, time, measurement, phone number, filename, date, price, percent, and ISBN. The overall average of the F-measure was 90.9%. Both of the previous two systems [44, 35] have evaluated their systems against ANERcorp.

4. NAMERAMA system

NAMERAMA is a named entity recognition system developed to extract named entities such as cancer disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic texts in the medical domain. The system consists of four main parts: pre-processing, data analysis, features extraction, and classification stage (Figure 1). At the pre-processing stage, data tokenisation and part-of speech (POS) tagging are carried out using the AMIRA tool. The data is checked and corrected manually after the tokenisation and POS tagging stages. Then, the data is annotated manually and each token in the data is given an appropriate semantic tag. At the data analysis stage, frequency analysis, collocation analysis and concordance analysis are carried out in order to extract the optimal features set. The Gazetteer features, lexical markers features, patterns features and a list of stopwords are used to support the POS tagging and data annotation. To extract the relevant entities, the system applies a probabilistic approach, a Bayesian Belief Network (BBN) which is a graphical representation of a probabilistic dependency. A BBN consists of a set of interconnected nodes, where each node represents a variable in the dependency model and the connecting arcs represent the causal relationships between these variables. Each node or variable may include a number of possible states/values. The belief in each of these states/values is determined from the belief in each possible state of every node directly connected to it and its relationship with each of these nodes. The belief in each state of a node is normally updated whenever the belief in each state of any directly connected node changes [48].

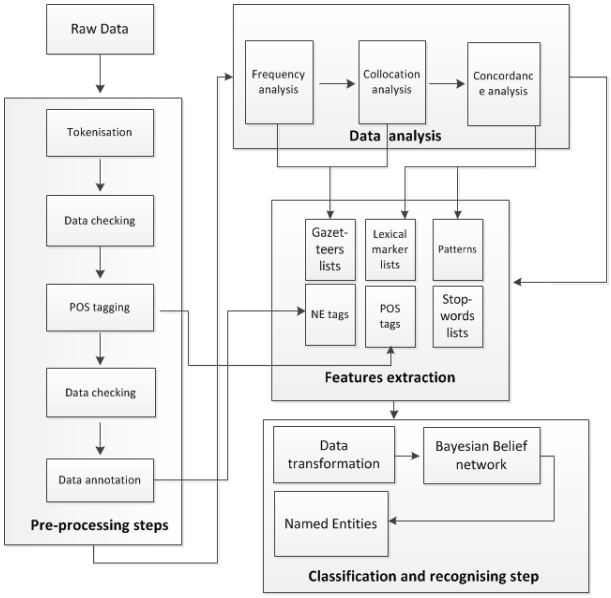


Fig 1: The NER system architecture.

4.1 Data description

NAMERAMA was tested on data obtained from the King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE) website. KAAHE was developed as a collaboration between the King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS) and the Saudi Association for Health Informatics (SAHI). The KAAHE website was later further developed in collaboration with the National Guard Health Affairs (NGHA), the Health on the Net Foundation (HON) and the World Health Organisation (WHO). KAAHE is regarded as a reliable health information source which contains information

about diseases, diet and healthy lifestyle, diagnosis and treatment, news and events, and Arabic medical resources, written in an easy and understandable language that fits with all community groups [9].. In this study, the initial focus is on cancer, and NAMERAMA was tested on a data set consisting of 26 articles with a total of 5119 tokens extracted from the diseases folder.

4.2 Pre-Processing phase

Tokenisation was carried out using AMIRA which is a toolkit consisting of a clitic tokeniser, part of speech tagger (POS) and base phrase chunker, using a shallow syntactic parser [18]. The tokenisation task may seem trivial in languages such as English, where a single space or punctuation is used to split sentences into words (tokens). In Arabic tokenization is a non-trivial algorithm due to the complex morphological structure of Arabic, and any errors made in this step can propagate into later phases and lead to serious problems. Clitic is a unit whose status lies in between that of an affix and a word. Some clitics can precede the word like a prefix and others follow the word like a suffix. For instance in the word وسيكتبونها "and they will write it" the conjunction "and" and the future marker "will" are represented as prefixes by the letterو andس respectively, while the pronouns "they" and "it" are represented by the suffixes ون and ها respectively. During the text tokenisation, the suffixes are not segmented because this increases the ambiguity and sparsity of the text, as there are more than 127 suffixes in Arabic [41]. The results of tokenization using AMIRA achieved 91%, 87% and 89% for precision, recall and F-measure, respectively. All identified errors were corrected manually before applying the following step in the pre-processing phase.

AMIRA was also used to perform POS tagging. Three different tag sets are provided: Bies Tag Set, Extended Reduced Tag set (ERTS), and Extended Reduced Tag set + Person information (ERTS_PER). ERTS was chosen for the POS tagging process because it has many important morphological features that are not included in the Bies tag set and Person information was a redundant feature as our data is written in the third person. AMIRA achieved 84% accuracy in POS tagging.

During the data annotation, four named entities (NEs) are recognised: disease name (D), symptoms (S), treatment methods (T) and diagnosis methods (G). Table 1 lists the numbers and the tags identified in this step.

Table 1: Numbers and the tags identified

Named entities	*Number of named entities*	*The tag*
Disease names	116	I_D
Symptoms	88	I_S
Treatment methods	79	I_T
Diagnosis methods	24	I_G

In the NER literature, the most used tagging schemes are the inside-outside (IO) and the inside-outside-beginning (IOB) schemes. In the IO tagging scheme, the tag I marks the word as being inside the named entity, and the tag O marks the word as being outside the named entity, while in the IOB tagging scheme the extra B tag marks the beginning of the named entity. Table 2 gives an example of how the sentence, "Salivary gland cancer is rare" is tagged by IO and IOB schemes.

Table 2: Tagging a sentence by IO and IOB schemes

Lexical items	*IO tagging*	*IOB tagging*
Salivary	I_ D	B_D
gland	I_D	I_D
cancer	I_D	I_D
is	O	O
rare	O	O

The IO tagging scheme is used in the data annotation step. Although this scheme cannot determine the boundary in the case of two named entities from the same class appearing next to each other, this does not affect the performance of this step as no such two named entities from the same class appear next to each other in our dataset. Furthermore, IO outperforms the IOB scheme in terms of cost and run time because it needs fewer tags in comparison with the IOB scheme. The number of tags in the IO scheme is (C +1) while in IOB it is (2C+1), where C is the number of named entity classes.

4.3 Data Analysis

This step involves three tasks: frequency analysis, collocation and concordance. Frequency analysis was used to compute the most frequent tokens in the data, using aConCorde 0.4.3, which is a multi-lingual concordance tool originally for Arabic concordance analysis [39]. Table 3 lists the 10 most frequent tokens. The most frequently occurring token was the conjunction "and" which occurred 276 times, and the least frequently occurring tokens with a count of 1 were words such as "مقدمة" "introduction" and "موقع" "location".

Table 3: The 10 most frequent tokens in our data.

Token	*Translation*	*Frequency*	*Token*	*Translation*	*Frequency*
و	And	276	ب	by, with, in	146
في	In	140	من	from	117
سرطان	cancer	113	أو	or	94
ل	to , for , so	68	المعالجة	the treatment	67
على	On	58	أن	that	55

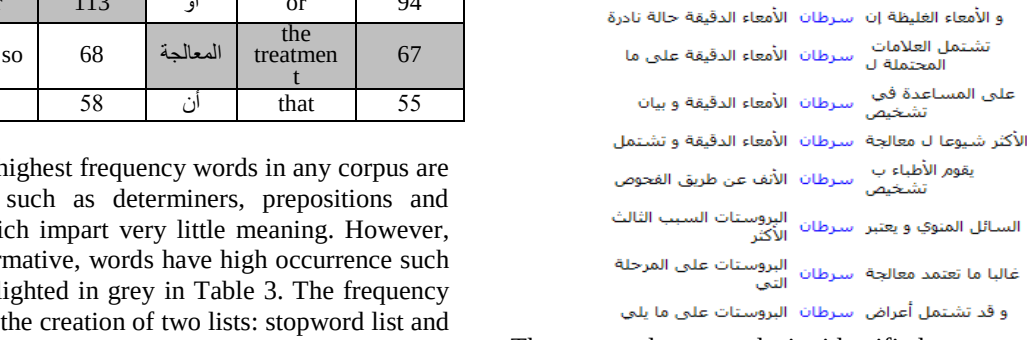
As expected the highest frequency words in any corpus are function words such as determiners, prepositions and conjunctions which impart very little meaning. However, other, more informative, words have high occurrence such as the ones highlighted in grey in Table 3. The frequency analysis enabled the creation of two lists: stopword list and gazetteers. Gazetteers are dictionaries that collect lists of relevant named entities. Four different gazetteers were created to capture the four important entities: disease names, symptoms, treatment methods and diagnosis methods. Collocation analysis was the next task. The term "collocation", which was first used by the linguist Firth [24] involves the combination of two words which are often used together. In our study, the collocation analysis is based on the text2ngram tool [51] which identified the 8 most frequently occurring informative collocations in our data (Table 4). As table 4 shows the Arabic collocation terms are listed with their embedded conjunctions and determiners.

Table 4: The 8 most frequent informative collocations in our data.

Collocation	*Translation*	*Frequency*
ب سرطان	by cancer	21
الإصابة ب	infection with	16
ل سرطان	of cancer	10
سرطان الشرج	anal cancer	9
المعالجة ب	treatment with	7
المعالجة الكيميائية	chemotherapy	18
المعالجة الشعاعية	radiation therapy	11
سرطان الرحم	cervical cancer	10

The collocation analysis enhances the gazetteers list by giving detailed information about certain entities. For instance, the collocation "المعالجة الشعاعية" is an entity related to the treatment method, the collocation " سرطان الشرج" and "سرطان الرحم" are entities representing disease names. The collocation analysis also produced the lexical marker lists–which are important lexical tokens used to indicate the presence or absence of named entity. For example, the words "الشعور" and "ألم" are lexical markers referring to the symptom entity.

The third task is concordance analysis which assists in the investigation of the context of named entity. It gives details about the structure of the language used in our medical domain. Furthermore, it leads to the identification of patterns in the data. The aConCorde 0.4.3 tool is used to carry out the concordance analysis. Figure 2 shows the concordance analysis for the word "سرطان".



The concordance analysis identified some verb-related patterns. These include the following verbs: "تتضمن" and "يشتمل" refer to the verb "include", "يحدث" to the verb "occur" and "تظهر" to the verb "appear". The pattern details are as follows:

Pattern 1: the verb "يتضمن" ←→ include
Structure: verb + entity + punctuation
Example 1: "تتضمن الأعراض:" ←→ The symptoms include:
Example 2: "تتضمن الخيارات العلاجية:" ←→ The treatment options include:

Pattern 2: the verb "يشتمل" ←→ involve
Structure: verb + named entity + preposition
Example: تشتمل سبل المعالجة على ←→ the treatment methods involve

Pattern 3: the verb "يحدث" ←→ occur
Structure: verb + disease name + disease type
Example: يحدث سرطان الرئة ←→ lung cancer occurs

Pattern 4: the verb "تظهر" ←→ appear
Structure: verb + symptoms + as + entity
Example: تظهر الأعراض على شكل كتلة في الثدي ←→ the symptoms appear as a lump in the breast

The concordance analysis also revealed some noun-based patterns. These include the following nouns: "تشخيص" "diagnosis" and "معالجة" "treatment". The patterns for these nouns are structured as follows:

Pattern 5: the noun "تشخيص" ←→ diagnosis
Structure: noun + by + diagnosis methods entity
Example: يتم التشخيص عن طريق الخزعة ←→ the diagnosis is by taking a biopsy

Pattern 6: the noun "معالجة" ←→ treatment
Structure: noun + cancer + cancer type
Example: تعتمد معالجة سرطان البروستات على مرحلته ←→ Prostate cancer treatment often depends on the stage of the cancer.

4.4 Feature extraction

Feature extraction is a crucial task in NER systems and is mostly machine learning based, in which the classifier depends entirely on the learning phase. Hence, selecting the optimal set of features enhances the performance of the classifier [14].
Based on the Arabic text analysis and the findings from the previous phase the following set of features was extracted from the corpus. This set wasbe used in the classification and recognition phase to train the Bayesian network. These features are:

- The POS tag of the word.
- Lists of lexical marker lists representing the entities.
- List of stopwords.
- Gazetteers for each entity type.
- The annotation tag for all words within a -/+2 word window.
- A set of patterns extracted from the study of the corpus.

4.5 Classification and recognition phase

Three main tasks are involved in this phase. First, the data is converted to a matrix where each row in the matrix is related to a token in the data, and each element in the row represents a feature (Figure 3)

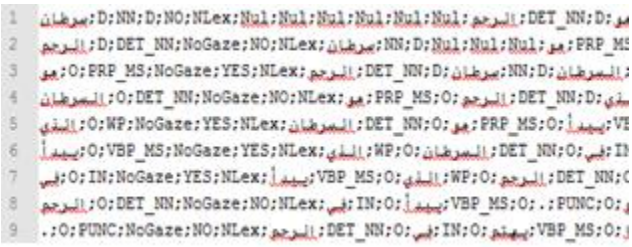


Fig. 3 A sample of the data after the transformation step

Then, a Bayesian Belief Network (BBN) is used to perform the classification and recognition task.

It allows us to form a hypothesis about the world based on observable variables (the given evidence e).

$$p(H|e) = p(e|H)p(H)\,p(e)$$

Where p(H|e) is sometimes called posterior probability, p(H) is called prior probability, p(e|H) is called likelihood of the evidence (data) and p(e) is just a normalising constant [26]. BBN is a directed graph, together with an associated set of probability tables. The graph consists of nodes and arcs where each node represents a feature, in the dependency model and the connecting arcs represent the causal relationships between these variables/features.

Our BBN is developed using the Bayes Server 6.7 tool. Figure 4 shows its structure which consists of 26 nodes and 25 arcs; each node is associated with a number of possible states/values representing its feature type. The belief in each of these states/values is determined from the belief in each possible state of every node directly connected to it and its relationship with each of these nodes. The target node, which needs to be predicted and recognised, is the entity node; it is dependent of the feature states of its neighbouring entities: $Entity_{i-1}$, $Entity_{i-2}$, $Entity_{i+1}$ and $Entity_{i+2}$. The predicted feature of the target entity is based on the following set of hypotheses:

- The feature node (e.g. lexical marker features, gazetteer features, pattern features, POS features, stopwords) values can lead to better prediction of the entity node.
- The entity types of the neighbouring two nodes of the target entity ($Entity_{i-1}$, $Entity_{i-2}$, $Entity_{i+1}$ and $Entity_{i+2}$) can lead to better prediction of the entity node.
- For a better prediction of the $Entity_{i-1}$, $Entity_{i-2}$, $Entity_{i+1}$ and $Entity_{i+2}$ nodes' values, their feature nodes need to be analysed against the set of patterns.

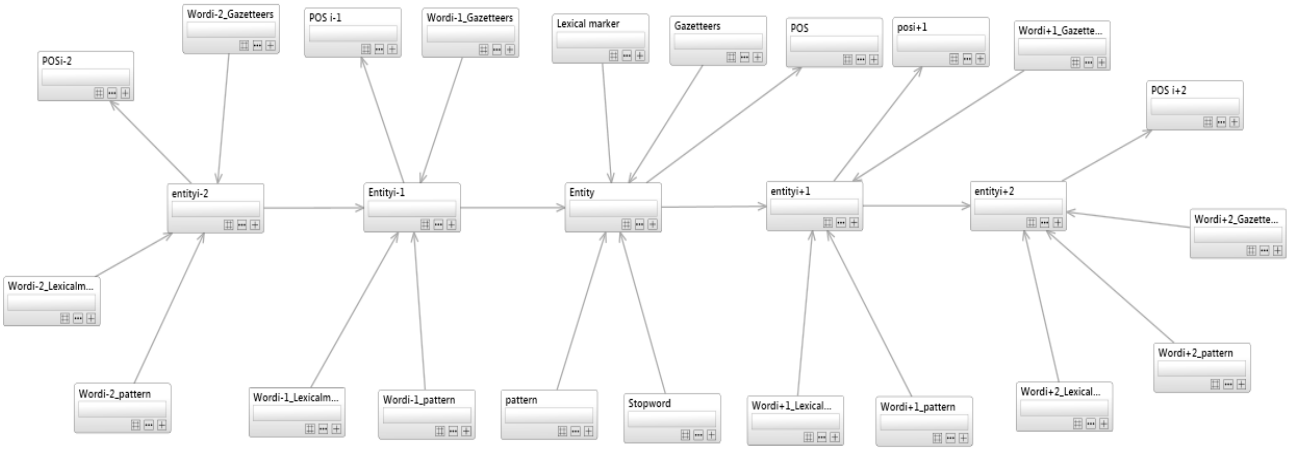


Fig. 4 The structure of our BBN.

The recognition involves a two steps process. First, the data is divided into training data, which constitutes 78%, and testing data, which constitutes 22 %. During the training phase, values of all nodes are provided to the BBN while during the testing phase the values of nodes Entityi-1, Entityi-2, Entityi+1 and Entityi+2 are predicted by calculating new beliefs based on new information or evidence.

5. Results and discussion

The Bayes server tool has a number of algorithms to perform inference for the prediction: relevance tree, likelihood sampling, loopy belief and variable elimination algorithms. The results of using these algorithms to recognise the entity were identical. Precision, recall, and F-measure metrics were used to evaluate the results. Table 5 lists the result for each entity type.

Table 5: the results of our BBN system.

Entity	*Precision*	*Recall*	*F-Measure*
Disease names	96.29%	100.00%	98.10%
Diagnosis methods	63.63%	41.17%	49.99%
Symptoms	68.18%	30.00%	41.66%
Treatment methods	55.73%	91.89%	69.38%
Overall	72.97%	69.23%	71.05%

The results show that the BBN approach is promising with an overall F-measure of 71.05%. The highest F-measure score was achieved in recognising disease names while the lowest was in recognising symptoms. Recognising disease names is straightforward compared to recognising symptoms. The data used comprised 26 articles related to different types of cancer. Therefore, recognising the different types of cancer was completed with a high recall and precision. On the other hand, symptoms are usually expressed in terms of long sentences. For instance, one of the symptoms in the data was " وجود كتلة في منطقة الأذن أو الوجنة أو الفك أو الشفة أو في داخل الفم" (lump in the area of the ear, cheek, jaw, lip or inside the mouth). This makes identifying the boundary of the symptom entity a challenging task. As a result, the F-measure score for entities related to symptom was lower than for other entities. The system performance in recognising the treatment methods entity was quite low at the precision metric compared to the recall. Although our approach correctly recognises most entities related to treatment methods , it incorrectly identified other tokens as treatment methods.

In the literature, there was no NER system for modern standard Arabic in the medical domain which supported the comparison of our results with other approaches. Current Arabic NER systems focus on recognising different sets of entities such as person, organisation and location, so comparing our results to theirs would not be helpful as our system uses different data and extracts different sets of entities. Instead, a baseline system was implemented to recognise the correct entities. The baseline system was based on gazetteers, so it automatically labels a token with an appropriate entity tag whenever this token is within a gazetteer file. Table 6 shows the results obtained from both systems in terms of F-measure.

Table 6: The obtained F-measures of the baseline and BBN systems.

Entity	*Baseline F-measure*	*BBN system F-measure*	*The difference*
Disease names	65.82%	98.10%	+ 32.28%
Diagnosis methods	36.35%	49.99%	+ 13.64%
Symptoms	17.85%	41.66%	+ 23.81%
Treatment methods	61.75%	69.38%	+ 7.63%
Overall	49.76%	71.05%	+ 21.29%

Table 6 shows clearly that our BBN approach outperforms the baseline system overall by 21%. The baseline system labels the majority of tokens as not entity because it relies on gazetteers only and our gazetteers have a limited number of tokens. As a result of this, the recall of the baseline system is very low (the symptoms entity achieves only 10% recall) while the precision is quite high. The results of the BBN approach can be improved by considering a number of factors. First, increasing the training data size could improve the training and the prediction of the entities. Second, choosing an optimal features set could increase system performance.

Acknowledgments

This research is supported by Aljouf University, Saudi Arabia and Staffordshire University, UK.

References

[1] Abdallah, S., Shaalan, K., and Shoaib, M., "Integrating rule-based system with classification for Arabic named entity recognition", in A Gelbukh, ed. Computational Linguistics and Intelligent Text Processing, Berlin Heidelberg, (7181), 2012 pp.11–322.

[2] AbdelRahman, S., Elarnaoty, M., Marwa M., and Fahmy, A., "Integrated machine learning techniques for Arabic named entity recognition",International Journal of Computer Science Issues (IJCSI) Vol. 7, 2010, pp.27–368.

[3] Abdul-Hamid, A. and Darwish, K., "Simplified feature set for Arabic named entity recognition", in Proceedings of the Named Entities Workshop, Stroudsburg, PA, 2010, pp.110–115.

[4] Aboaoga M., and Aziz M, "Arabic person names recognition by using a rule based approach", Journal of Computer Science, (9), 2013, pp.922–927.
[5] Algahtani, S., "Arabic Named Entity Recognition: A Corpus-Based Study". Ph.D. thesis, The University of Manchester, UK, 2011.
[6] Al-Jumaily, H., Paloma,M., Jos, M., and Erik G., "A real time named entity recognition system for Arabic text mining", Language Resources and Evaluation Journal, Vol. 46(4), 2012, pp.543–563.
[7] Alkharashi, I., "Person named entity generation and recognition for Arabic language", In Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, 2009, pp.205–208.
[8] Al-Shalabi, R., Ghassan K., Al-Sarayreh, B., Khanfar, K., AIGhonmein, A. Talhouni, H., and Al-Azazmeh, S. (2009) Proper noun extracting algorithm for the Arabic language. In: International Conference on IT to Celebrate S. Charmonman's 72nd Birthday, Bangkok, pp.28.1–28.9.
[9] Alsughayr A., "King Abdullah Bin Abdulaziz Arabic health encyclopedia (www.kaahe.org): A reliable source for health information in Arabic in the internet", Saudi J Med Med Sci; Vol. 1: 53, 2013, pp. 4.
[10] Althobaiti, M., Kruschwitz, U., and Poesio, M., "A Semi-supervised Learning Approach to Arabic Named Entity Recognition" University of Essex, UK, 2013.
[11] Asharef, M., Omar, N., and ALBARED, M., "Arabic Named Entity Recognition in Crime Documents", Journal of Theoretical & Applied Information Technology, Vol. 44, 2012, pp. 1-6.
[12] Benajiba, Y., Diab, D., and Paolo R., "Arabic named entity recognition: A feature-driven study", in IEEE Transactions on Audio, Speech, and Language Processing, 2009, pp.926–934.
[13] Benajiba, Y., and Paolo, R., "Arabic named entity recognition using conditional random fields", In Proceedings of the Workshop on HLT & NLP within the Sixth International Conference on Language Resources and Evaluation (LREC), 2008, Marrakech, pp.143–153.
[14] Benajiba, Y., Diab. M, and Paolo R., "Arabic named entity recognition: An SVM-based approach", in Proceedings of Arab International Conference on Information Technology (ACIT), 2008, Hammamet, pp.16–18.
[15] Benajiba, Y., and Paolo, R., "ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information", in Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI), 2007, Mumbay, pp.1814–1823.
[16] Bidhend, M., Behrouz M., and Hosein J., "Extracting person names from ancient Islamic Arabic texts", in Proceedings of Language Resources and Evaluation for Religious Texts (LRE-Rel) Workshop Programme, Eight International conference on Language Resources and Evaluation (LREC), 2012, Istanbul, pp.1–6.
[17] Bodnari, A., Deleger, L., Lavergne, T., Neveol, A., and Zweigenbaum, P., "A supervised named-entity extraction system for medical text", In Online Working Notes of the CLEF, 2013, Evaluation Labs and Workshop, September.
[18] Diab, M., "Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging and Base Phrase Chunking", in Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009.
[19] Di Pietro, G., Aliprandi, C., De Luca, A. E., Raffaelli, M., and Soru, T., "Semantic crawling: An approach based on Named Entity Recognition", in Advances in Social Networks Analysis and Mining (ASONAM), 2014, IEEE/ACM International Conference on (pp. 695-699).
[20] Elsebai, A., "A Rules Based System for Named Entity Recognition in Modern Standard Arabic", PhD thesis, University of Salford, UK, 2009.
[21] Eltyeb, S., Salim, N., "Chemical named entities recognition: a review on approaches and applications", Journal of Cheminformatics, Vol. 6:17, 2014.
[22] Farber, B., Dayne F., Habash,H. and Owen, R, "Improving NER in Arabic using a morphological tagger", in proceedings of the Sixth International Conference on Language Resources and Evaluation. (LREC), 2008, Marrakech , pages 2,509–2,514.
[23] Farghaly, A., and Shaalan, K., "Arabic natural language processing: Challenges and solutions", in ACM Transactions on Asian Language Information Processing (TALIP), 2009, pp.1–22.
[24] Firth, J., Papers in Linguistics 1934-1951. Oxford: Oxford University Press, 1957.
[25] Guo, J., Gu X., Xueqi C., and Hang Li., "Named entity recognition in query", In Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2009, New York City, pp.267–274.
[26] Heckerman, D, A tutorial on learning with Bayesian networks. Springer Netherlands.
[27] Kothari, C., Research Methodology; Methods and Technique Dharmesh Printers. New Delhi, India, 2004.
[28] Koulali, R, and Abdelouafi, M., "A contribution to Arabic named entity recognition", In Proceedings of 10th International Conference on ICT and Knowledge Engineering, 2012, Morocco, pp.46–52.
[29] Maloney, J., and 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, Stroudsburg, PA, pp.8–15.
[30] Mohammed. N, and Omar, N., "Arabic Named Entity Recognition Using Artificial Neural Network", Journal of Computer Science, pp.1285-1293, 2012.
[31] Nadeau, D. and Sekine S., "A survey of named entity recognition and classification", Lingvisticae Investigationes, pp.3–26, 2007.
[32] Nadeau, D., and Sekine, S. "A survey of named entity recognition and classification", In Satoshi Sekine(Editor), Elisabete Ranchhod(Editor)Named Entities: Recognition, classification and use (Benjamins Current Topics) Hardcover– 3 Jul 2009
[33] Nezda, L, Andrew H, John L, and Sarmad F., "What in the world is a shahab? Wide coverage named entity recognition for Arabic", in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), 2006, Genoa, pp.41–46.

[34] O'Steen, D., and Breeden, D., "Named Entity Recognition in Arabic: A Combined Approach" BA (Hons), Stanford University, USA, 2009.
[35] Oudah, M., and Shaalan, K., "Person name recognition using the hybrid approach", Natural Language Processing and Information Systems, Berlin Heidelberg, (7934), pp.237–248, 2013.
[36] Petasis, G., Cucchiarelli, A., Velardi, P., Paliouras, G., Karkaletsis, V., and Spyropoulos, C., "Automatic adaptation of Proper Noun Dictionaries through cooperation of machine learning and probabilistic methods", In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 128-135.
[37] Rau, L., "Extracting Company Names from Text", in Proc. Conference on Artificial Intelligence Applications of IEEE, 1991.
[38] Rizzo, G., Marieke, v., and Raphaël, T., "Benchmarking the extraction and disambiguation of named entities on the semantic web", In 9th International Conference on Language Resources and Evaluation (LREC), 2014.
[39] Roberts, A.; Al-Sulaiti, L., and Atwell, E., "aConCorde: towards a proper concordance of Arabic in", in Proceedings of Corpus Linguistics 2005.
[40] Samy, D., Moreno-Sandoval, A., Bueno-Diaz, C., Garrote-Salazr, M., and Guirao, J., "Medical Term Extraction in an Arabic Medical Corpus", in Proceedings of the 8th Language Resources and Evaluation Conference, 2012, Istanbul, Turkey.
[41] Sawalha, M. and Atwell, E., "Linguistically Informed and Corpus Informed Morphological Analysis of Arabic", In Proceedings of the 5th International Corpus Linguistics Conference CL, 20-23 July 2009, Liverpool, UK.
[42] Shaalan, K., "A Survey of Arabic Named Entity Recognition and Classification", Computational Linguistics, 40:2, MIT Press, 2014.
[43] Shaalan, K., "Rule-based Approach in Arabic Natural Language Processing", The International Journal on Information and Communication Technologies (IJICT), 2010, pp.11-19.
[44] Shaalan, K., and Raza, H., "Arabic named entity recognition from diverse text types", in Advances in Natural Language Processing, 2008, Vol. 5221 , pp.440–451.
[45] Shihadeh, C., and Neumann, G., "ARNE: A tool for named entity recognition from Arabic text", In Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4), 2012, San Diego, CA, pp.24–31.
[46] Sundheim, B., "Overview of results of the MUC-6 evaluation", Proceedings of a workshop on held at Vienna, 6-8 May 1996, pp.423-442.
[47] Traboulsi, H., "Arabic named entity extraction: A local grammar-based approach", in Proceedings of the International Multi-conference on Computer Science and Information Technology (IMCSIT),2009, Mragowo, pp.139–143.
[48] Wooldridge, S., Bayesian Belief Networks. Centre for Complex System Science, CSIRO, Canberra, 2003.
[49] Zaghouani, W., Pouliquen, B., Ebrahim, M., and Steinberger, R., "Adapting a resource-light highly multilingual named entity recognition system to Arabic", in Proceedings of the Seventh International Conference on Language Resources and Evaluation, 2010, Valletta, pp.563–567.
[50] Zaghouani,W., "RENAR: A rule-based Arabic named entity recognition system" ACMTransactions on Asian Language Information Processing (TALIP), 2012, 11(1):2:1–2:13.
[51] Zhang, L.: Text2Ngram. http://homepages.inf.ed.ac.uk/s0450736 /ngram.html
[52] Zayed, O., and El-Beltagy, S., "Person name extraction from modern standard Arabic or colloquial text", In Proceedings of the 8th International Conference on Informatics and Systems conference, 2012, Cairo, pp.44–48.

Saad Alanazi received his B.S degree in Computer Science from Aljouf University in 2007 and M.S. degree in Computer Science from Ball State University in 2011. He is currently a Ph.D student at Staffordshire University. His research interests include natural language processing and text mining.

Bernadette Sharp is Professor of Applied AI at Staffordshire University. She is a Chartered IT Professional Fellow of the British Computer Society. She has published over 100 referred publications in the areas of applied artificial intelligence, natural language processing and knowledge discovery. She is Chair and editor of the International Workshop for Natural Language Processing and Cognitive Science (NLPCS) and the International Conference on Agents and Artificial Intelligence between 2009-2010. She has BSc in Computer Mathematics, MPhil in Statistical Forecasting, PhD in Natural Language Processing.

Clare Stanier is a Senior Lecturer in Information Systems at Staffordshire University. She is a Senior Fellow of the Higher Education Academy, a member of the British Computer Society and a programme committee member for TLAD, the HEA sponsored international workshop on the Teaching, Learning and Assessment of Databases. Her research interests are in data management and in Big Data strategies and technologies. She has an MSc in Business Intelligence and a PhD in Computer Science.