

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

SAAD ALANAZI

A thesis submitted in partial fulfilment of the requirements of Staffordshire University for the
degree of Doctor of Philosophy

May 2017

Abstract

Currently, 30-35% of the global population uses the Internet. Furthermore, there is a rapidly increasing number of non-English language internet users, accompanied by an also increasing amount of unstructured text online. One area replete with underexploited online text is the Arabic medical domain, and one method that can be used to extract valuable data from Arabic medical texts is Named Entity Recognition (NER). NER is the process by which a system can automatically detect and categorise Named Entities (NE). NER has numerous applications in many domains, and medical texts are no exception. NER applied to the medical domain could assist in detection of patterns in medical records, allowing doctors to make better diagnoses and treatment decisions, enabling medical staff to quickly assess a patient's records and ensuring that patients are informed about their data, as just a few examples. However, all these applications would require a very high level of accuracy. To improve the accuracy of NER in this domain, new approaches need to be developed that are tailored to the types of named entities to be extracted and categorised.

In an effort to solve this problem, this research applied Bayesian Belief Networks (BBN) to the process. BBN, a probabilistic model for prediction of random variables and their dependencies, can be used to detect and predict entities. The aim of this research is to apply BBN to the NER task to extract relevant medical entities such as disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic texts in the medical domain. To achieve this aim, a new corpus related to the medical domain has been built and annotated. Our BBN approach achieved a 96.60% precision, 90.79% recall, and 93.60% F-measure for the disease entity, while for the treatment method entity, it achieved 69.33%, 70.99%, and 70.15% for precision, recall, and F-measure, respectively. For the diagnosis method and symptom categories, our system achieved 84.91% and 71.34%, respectively, for precision, 53.36% and 49.34%, respectively, for recall, and 65.53% and 58.33%, for F-measure, respectively. Our BBN strategy achieved good accuracy for NEs in the categories of disease and treatment method. However, the average word length of the other two NE categories observed, diagnosis method and symptom, may have had a negative effect on their accuracy. Overall, the application of BBN to Arabic medical NER is successful, but more development is needed to improve accuracy to a standard at which the results can be applied to real medical systems.

Acknowledgements

There are many people without whom this work would not have been possible. I would firstly like to thank my supervisors Prof. Sharp Bernadette and Dr Clare Stanier who worked tirelessly with me during the development of my thesis. Their insight and ideas were an endless source of inspiration to me, driving my research at once to greater detail and to greater strength.

I would also like to give thanks to the Government of Saudi Arabia represented by Al-Jouf University for their financial support during my Ph.D. journey.

Finally, special thanks also to my parents, my wife, my son, and my brothers and sisters for their support and prayers. Words cannot express how grateful I am to have you in my life.

Publications

Alanazi, S., Sharp, B., & Stanier, C. (2015). An Evaluation of AMIRA for Named Entity Recognition in Arabic Medical Texts. *Natural Language Processing and Cognitive Science* , p.21-30.

Alanazi, S., Sharp, B., & Stanier, C. (2015). A Named Entity Recognition System Applied to Arabic Text in the Medical Domain. *International Journal of Computer Science Issues (IJCSI)*, 12(3), p.109-117.

Table of Contents

1	Chapter 1: Introduction	1
1.1	Background.....	1
1.2	Research Motivation.....	1
1.3	Aims and Objectives.....	3
1.4	Research Contributions.....	4
1.5	Research Methodology	5
1.6	Research Design	8
1.7	Challenges of Arabic Language Processing	9
1.8	Ethical Issues	10
1.9	Thesis Structure	11
2	Chapter 2: Literature Survey.....	12
2.1	Introduction	12
2.2	Evaluation Metrics of NER	12
2.3	Applications of NER	13
2.4	NER Approaches	16
2.4.1	Rule-based approach (the hand-written approach).....	23
2.4.2	Machine-learning approach.....	25
2.4.3	Hybrid approach.....	31
2.5	Feature Space for NER.....	32
2.6	NER Tools and Resources for Arabic Language.....	34
2.6.1	Corpora	34
2.6.2	Morphological analysers	37
2.7	Conclusion	39
3	Chapter 3: Theoretical Foundations.....	41
3.1	Introduction	41
3.2	Natural Language Processing Levels.....	41

3.3	Bayesian Belief Network.....	46
3.4	Conclusion.....	50
4	Chapter 4: Named Entity Recognition with NAMERAMA	52
4.1	Introduction	52
4.2	System Overview of NAMERAMA.....	52
4.3	Corpus Description	53
4.4	Arabic Language Processing Component.....	53
4.4.1	Data pre-processing steps.....	53
4.4.2	Data Analysis step.....	72
4.4.3	Features Extraction step.....	83
4.5	Summary.....	99
5	Chapter 5: Bayesian Named Entity Network	100
5.1	Introduction	100
5.2	Data Transformation Step.....	100
5.3	Feature Ranking.....	101
5.3.1	Likelihood-ratio	101
5.3.2	Naïve Bayes network	102
5.4	Naïve Bayes Network application to named entities	107
5.4.1	Evaluation of the NBN classifier	107
5.4.2	Optimising the Naïve Bayes Network performance	108
5.5	Bayesian Belief Network.....	112
5.5.1	BBN structure	113
5.5.2	Evaluation of the BBN classifier	114
5.5.3	Errors analysis.....	115
5.5	The Effect of the Sliding Window Size on the BBN Performance.....	123
5.6	Five-fold Cross-validation.....	128
5.7	Summary.....	130
6	Chapter 6: Conclusion.....	132

6.1	Review of the study	132
6.2	Complexity of Arabic Language and Limitations	135
6.3	Novel Contributions	136
6.4	Conclusion and Future Directions	138
7	References list.....	140
8	Appendix I. Stopwords list.....	154
9	Appendix II. Gazetteers	155
10	Appendix III. Lexical Markers.....	156

List of Figures

Figure 1.1 The research 'onion' (Saunders et al., 2009).	6
Figure 1.2 The main steps in quantitative research (Bryman, 2004).	8
Figure 1.3 The main steps in our research	9
Figure 2.1 NER approaches.	16
Figure 2.2 A visualisation aid of precision and recall.	35
Figure 3.1 Theories that underpin this research.	41
Figure 3.2 An example of a typical Bayesian network.	47
Figure 4.1 The architecture of NAMERAMA system.	54
Figure 4.2 A sample of the un-tokenised data in Arabic, its translation in English, and its tokenised version.	55
Figure 4.3 A sample of the tokenisation task result.	58
Figure 4.4 A sample of the POS tagging task result.	62
Figure 4.5 A sample of the annotated data.	67
Figure 4.6 The concordance analysis for the lexical item 'cancer', 'سرطان'.	80
Figure 4.7 the sketch of the word سرطان "cancer".	82
Figure 4.8 the result of clicking on the first verb right, تشمل "include"	83
Figure 4.9 The POS tags of a sample of the corpus in which the highlighted words are named entities of the category "disease names".	84
Figure 4.10 The concordance lines of the lexical item diagnosis "تشخيص"	85
Figure 4.11 Some examples of disease NE genitive nouns.	87
Figure 4.12 The verb-related patterns.	90
Figure 4.13 The disease NE noun-related patterns.	91
Figure 4.14 The symptom NE noun-related patterns.	96
Figure 4.15 The treatment methods NE noun-related patterns.	96
Figure 4.16 The diagnosis method NE noun-related patterns.	98
Figure 5.1 A typical Naïve Bayes network (Ang et al., 2016).	102
Figure 5.2 A simple Naïve Bayes network to evaluate the features.	103
Figure 5.3 The structure of our NBN.	110
Figure 5.4 The Venn diagram representation of the NBN performance.	111
Figure 5.5 The structure of the optimised NBN.	112
Figure 5.6 An example of a GBN (Ang et al., 2016).	113
Figure 5.7 The structure of our BBN.	116

Figure 5.8 The Venn diagram representation of the BBN performance.	117
Figure 5.9 A sample of the output of our BBN when recognising the disease entity.	118
Figure 5.10 A sample of the output of our BBN when recognising the treatment method entity.....	120
Figure 5.11 A sample of the output of our BBN when recognising the symptom entity.....	122
Figure 5.12 A sample of the output of our BBN when recognising the diagnosis method entity.....	123
Figure 5.13 Visualisation of different sliding window sizes.....	124
Figure 6.1 F-measure of BBN and baseline system.	134

List of Tables

Table 2.1 A Summary of the related Arabic NER work.	17
Table 2.2 The most common features (Shaalán, 2014).....	33
Table 4.1 Data size before and after the tokenising stage.....	56
Table 4.2 The main tokenisation schemes of AMIRA.....	57
Table 4.3 List of MADAMIRA POS Tags.	65
Table 4.4 IO and IOB schemes	66
Table 4.5 Total number of entities and tokens in our corpus.....	66
Table 4.6 Disease entities and tokens in the corpus.....	68
Table 4.7 Number of symptoms entities and tokens extracted.	69
Table 4.8 Number of treatment entities and tokens extracted.....	70
Table 4.9 Number of diagnostic entities and tokens extracted.	71
Table 4.10 The frequency distribution of the lexical items in the corpus.	74
Table 4.11 The 30 most frequent lexical items in our corpus.	75
Table 4.12 The frequency distribution of the collocations among our corpus.....	77
Table 4.13 The ten most recurrent collocations in our data.	77
Table 4.14 The most twenty informative collocations in terms of their frequency among our corpus. .	78
Table 5.1 The results of applying the NBN using each feature for the category of disease entities.....	104
Table 5.2 The results of applying the NBN using each feature for the diagnosis method entity category.	105
Table 5.3 The results of applying the NBN using each feature for the treatment method entity category.	106
Table 5.4 The results of applying the NBN using each feature for the symptom entity category.	107
Table 5.5 The results from our NBN.	108
Table 5.6 The results of our NBN using two different sets of features: all features and only the highest ranked feature.....	109
Table 5.7 The results from the optimised NBN.	109
Table 5.8 The performance of the NBN depending on the number of features used.....	112
Table 5.9 The results of our BBN network.	114
Table 5.10 Comparative analysis of the results achieved by the BBN and three NBNs.....	115
Table 5.11 Error results for the disease entity (D).	118
Table 5.12 Error results for the treatment method entity (T).	119

Table 5.13 Error results for the symptom entity (S).	121
Table 5.14 Error results for the diagnosis method entity (G).	121
Table 5.15 The results of our BBN network per sliding window size for the disease entity.....	125
Table 5.16 The results of our BBN network per sliding window size for the diagnosis method entity.	126
Table 5.17 The results of our BBN network per sliding window size for the treatment methods entity.	127
Table 5.18 The results of our BBN network per sliding window size for the symptom entity.....	127
Table 5.19 Disease entity: five rounds experiment.....	128
Table 5.20 Diagnosis method entity: five rounds experiment.	129
Table 5.21 Treatment method entity: five rounds experiment.	129
Table 5.22 Symptom entity: five rounds experiment.....	130
Table 6.1 The results of the BBN and baseline systems.	134

Abbreviations List

Abbreviations	Full form
ACE	Automatic Content Extraction
BBN	Bayesian Belief Networks
CoNLL	The Conference on Computational Natural Language Learning
CRFs	Conditional Random Fields
D	Disease Entity
DBN	Dynamic Bayesian network
G	Diagnosis method Entity
IO	Inside-Outside
IOB	Inside-Outside-Beginning
ME	Maximum Entropy
ML	Machine Learning
MSA	Modern Standard Arabic
MUC	Message Understanding Conference
NBN	Naïve Bayesian Network
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
POS	Part of Speech
S	Symptom Entity
SL	Supervised learning
SSL	Semi-Supervised Learning

SVM	Support Vector Machines
T	Treatment method Entity

Chapter 1: Introduction

1.1 Background

Named Entity Recognition (NER) was introduced at the sixth Message Understanding Conference (MUC-6) in 1995, which aimed at developing new methods to enhance the information extraction process. NER is a sub-area of information extraction whose goal is to extract specific predefined list of entities, which can include proper names, numerical expression and temporal expression. Each entity type can be divided into subtypes such as person names, location names, organisation names which can be classified as proper names, and money and percentages which fall under numerical expression type, whereas date and time are considered temporal expressions (Sundheim, 1996).

NER is essential for many natural language processing applications such as question answering, machine translation and information retrieval. Question answering systems usually provide an answer to questions such as who, what, when and where in the form of named entities. Therefore, developing an accurate named entity recognition system will help question answering systems to extract the correct answer. Moreover, named entities play a significant role in assisting the machine translation process as many named entities could be normal nouns, for example, the word "Apple" could be a fruit or a company name, although as a company name it would be capitalised. An accurate NER system should distinguish between these ambiguous meanings. Thus, a successful machine translation system should assign the appropriate meaning to words (Shalan, 2014). Also, NER can improve the information retrieval process as Guo (2009) indicated that more than 70% of research engine enquiries contain named entities.

1.2 Research Motivation

Named Entity Recognition is not only key to the creation and development of information extraction and management applications necessary to the modern world but also to text mining. Named Entity Recognition needs to be particularly fine-tuned for textual mining tasks, as it needs to extract all relevant information from particular categories so as to construct a reliable database or to discover associations among entities. For these applications to be valuable, it is necessary for the textual data to be well managed, and efficiently and effectively structured. Once information has been extracted it needs to be ordered into a suitable database for later tasks. Hale (2005) observed that most existing textual data used in the context of medicine is unstructured, with free text literature comprising 80% of all available data. Larger corpora will reveal patterns which make database construction and

organisation more effective by mining the token span of each located entity, identifying the structures of the entity, labelling each structure type and inferring their relationships. This allows the researcher to identify semantically rich structures and to conceptually summarise the data. Once structures have been identified, isolated and labelled, they can then be used for data retrieval and more complex tasks such as knowledge discovery. This is particularly important for real-world data, which rarely shows its structure in small-scale samples, requiring large corpora analysis to extract enough information to develop an understanding of the structure and context of particular named entities. For instance, in the medical domain, many clinical records appear in the form of narrative text, which has been dictated and later transcribed, or entered directly using informal and concise structure. Based on Hale's observations, we can note that health records, electronic medical records, case reports, patents, and news distribution such as research articles, news reports, blogs and even forums are not structured. This shows the importance of developing a strong named entity recognition system to support the medical community; such a system would assist in identifying named entities and categorising their surrounding structures.

An area of the medical domain which has not been adequately researched in Arabic is medical texts pertaining to cancer. As with most modern nations, Saudi Arabia suffers from ever increasing cancer diagnoses. In 2002, between the 1st of January and the 31st of December, there were a total of 7,942 cases of cancer reported to the Saudi Cancer Registry. In 2012, a mere ten years later, the reported cases had almost doubled, at 14,846 total cases. This is such a significant leap that it cannot be ignored, regardless of whether the increase is caused by an increase in cancer or an increase in diagnosis. It is clear that more research is needed to better understand the sudden increase in reported cases, so that appropriate action can be taken by the medical community. Furthermore, there was a demographic shift in that decade. In 2002, cancer was more commonly reported among men than women, with men representing 51.48% of cancer cases, and women representing 48.6%. However, in 2012 women were diagnosed with more cancer than men, with 52.5% of cases affecting women and 47.5% of cases affecting men. The overall age-standardised reported incidence rate for cancer had therefore leapt by almost 20 points, from around 60 per hundred thousand for women and sixty-two per hundred thousand for men, to seventy-eight per hundred thousand for men and over eighty-six per hundred thousand for women. The reasons for the difference between men and women, as well as women's sudden increase in reported cancer rates, need to be understood.

The current body of research and analysis conducted in the Arabic-language medical domain is insufficient, and, despite an obvious need for it, no research has been conducted in the field of cancer despite the alarming increase in reported cancer incidents in Saudi Arabia. Named Entity Recognition would assist in extracting valuable patterns from all available reports and constructing a database from which the new cases can be better understood. This has motivated this study to focus on extracting

named entities from Arabic medical corpus as it can contribute towards advancing cancer patient care in Saudi Arabia.

There are many potential advantages to developing a reliable Named Entity Recognition system for the medical domain. These advantages include:

- The creation of systems which could summarise unstructured medical text to provide key, relevant information on reports, blogs or websites for the patient to access.
- Simplifying the process of cancer diagnosis by applying targeted Named Entity Recognition so as to extract data which can be used for determining similarities between various cases. This could connect raw data, list laboratory test medications and outcomes by frequency or by treatment method, allowing for a sounder decision to be made effectively and in a timely fashion.
- Access to patient information and automatic as opposed to manual retrieval of relevant data could allow the patient to be treated sooner and more effectively.
- The creation of user friendly systems which could explain jargon and technical terms or translate them into simple language, so as to make reports more easily understood by patients, families, non-native speakers of Arabic, doctors from other medical specialties or legal representatives.
- The tracking of the medical history of each patient across hospitals and the delivery of relevant information for outpatients via named entities could help the sharing of best practices among medical staff.
- The creation of computerised decision support systems would ensure not only fast provision of services and ensure compliance with best clinical practices, but also enrich existing databases and literature-based knowledge bases, as well as discover new knowledge and useful trends.

1.3 Aims and Objectives

The aim of this research is to develop a novel BBN based Named Entity Recognition (NER) approach to extract relevant medical entities such as disease names, symptoms, treatment methods, and diagnosis methods from modern Arabic texts in the medical domain. This research aims also to answer the following research questions. How does the use of Bayesian belief network contribute in the context of named entity recognition task for Arabic text? How supportive is it?

Bayesian networks are an increasingly popular method of modelling uncertain and complex domains. They have been deployed successfully in many applications including the healthcare and medical domains. They have also been successfully used for English language processing, such as spell

checking (Haug et al., 2001), text categorisation and retrieval (Yang, 1994), and speech recognition (Bilmes, 2004).

Bayesian Belief Networks (BBN) focus on the probability of an event E , to capture the degree of belief or confidence of the occurrence of E based on prior and observed facts or knowledge. They have several advantages for data analysis over rule based or decision trees. The belief can be updated when new data arrives. They allow learning to take place and can improve prediction based on observations. They can deal with incomplete data and they are particularly suitable to represent medical knowledge and for diagnostic purposes. For example, they have been successful at representing probabilistic relationships between diseases and symptoms. They have been used by the machine learning community to compute the probabilities of the presence of various diseases based on a given set of symptoms. Another advantage of Bayesian networks is that it is intuitive and close to human's strategy in understanding observed facts and their impact or effects on other relations. As our aim is to extract entities based on observed linguistic descriptive features and our application relates to the medical domain we wish to investigate whether BBN can learn from these observed features and patterns, can extract the desired named entities and update its learning as more features or patterns are identified.

The aims of this study can be achieved by fulfilling the following objectives:

- To survey the current Arabic NER systems and methodologies.
- To build and annotate a corpus related to the medical domain extracted from a well-respected and widely used website such as the King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE) website.
- To develop a novel approach to NER based on Bayesian Belief Network (BBN).
- To implement and test the novel approach to the above annotated corpus using an appropriate methodology.
- To evaluate the novel approach using well-known measures, such as precision, recall, and F-measure and to evaluate the effectiveness of the use of Bayesian belief networks in the context of NER for Arabic text
- To validate our approach using k-fold cross validation approach.

1.4 Research Contributions

The major novel contributions of this research project include the following:

- The application of BBN to the named entity recognition task.
- The application of BBN to analyse modern standard Arabic (MSA) texts.

- The application of BBN to the extraction of complex medical entities.
- The production of a manually annotated medical corpus in MSA.

Other minor contributions are listed below:

- The evaluation of AMIRA tool performance in terms of the tokenisation and part of speech (POS) tagging processes.
- Measuring the impact of using different features alongside BBN on the NER task.
- Assessing the impact of using different sliding window sizes on the performance of the NER task.

1.5 Research Methodology

A well-known way to explain the research philosophy, approach and strategy is using the research onion. Saunders *et al.* (2009) outline the concept of the “research onion” (Figure 1.1), which refers to the layers of research which need to be addressed before matters such as data collection can begin. They divide the research onion's layers, from outermost to innermost, into research philosophy, research approach, research strategy, choices, time horizons, and finally conclude with data collection techniques and data analysis procedures.

In this section we will briefly address the first three layers of the onion, research philosophy, research approach and research strategy as applied to our research.

- Research philosophy

Saunders *et al.* (2009) explain that research philosophy covers the development of knowledge and an understanding of its nature within your field of research. The research philosophy chosen will establish how the research strategy is developed as well as which methods may be employed under that strategy. This means that different research philosophies may be more or less conducive to the retrieval and analysis of accurate data depending on the field of research and the question at hand. Saunders *et al.* describe the four main research philosophies as Positivism, Realism, Interpretivism and Pragmatism.

This research adopts Positivism which is also known as “scientific philosophy”. According to Easterbay-Smith *et al.* (2012) “the key idea of positivism is that the social world exists externally, and that its properties should be measured through objective methods” (Easterbay-Smith *et al.*, 2012, p.22). Under Positivism the researcher will only handle observable phenomena and will include a process of operationalisation, by which the reality we are observing is translated into measurable figures. Our research is based on currently available, reliable data, allowing us to work with observable phenomena

and therefore lending itself to Positivist philosophy. We will be applying verifiable algorithms and our results will be capable of being replicated.

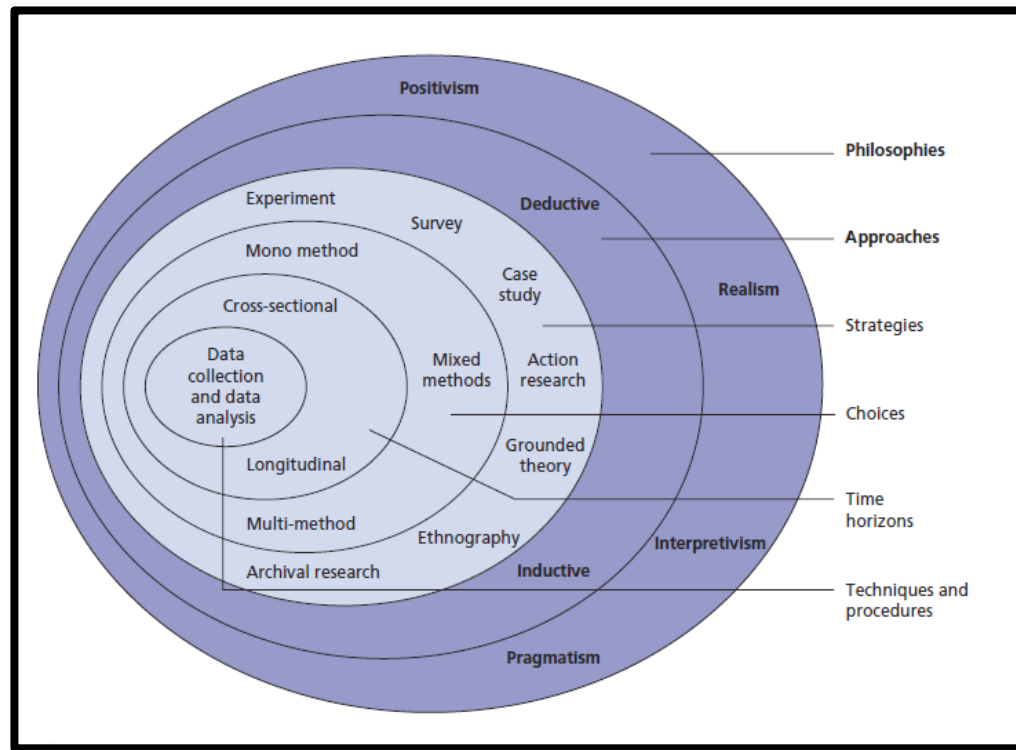


Figure 1.1 The research 'onion' (Saunders et al., 2009).

- Research approach

The research approach to be adopted for this study is deductive. Deductive research follows the development of a hypothesis which is then submitted to thorough testing. This approach is very common across all objectively scientific research and is essential to a Positivist research philosophy, as well as to quantitative research in general.

According to Robson (2002) there are five essential stages for deductive research, which have parallels in Bryman's steps of quantitative research, discussed in 1.6. Robson lists the deduction of a hypothesis from the original theory as the first step. This corresponds with Bryman's steps one and two. Robson lists expressing the hypothesis in operational terms as the second step. This corresponds with Bryman's steps three and four. Robson then lists the test of this hypothesis as the third step. This corresponds with Bryman's steps five, six and seven. Robson lists the examination of the outcomes of the test as the fourth step. This corresponds with Bryman's steps eight and nine. The fifth and final step in Robson's list is the modification of the original theory, as required, to account for the results of the test. This corresponds to Bryman's steps ten and eleven. The close connection between the steps of deductive

research and the steps of quantitative research in general shows the strengths of deductive research to a positivist researcher.

The merits of applying a deductive research approach are many. Deduction supports establishing reasons for casual relationships between variables, it lends itself particularly well to the handling of raw data and the operationalisation of observable phenomena (Saunders *et al.*, 2009). On the other hand, an inductive approach is designed to handle the interpretation of data whereby on its own the relationship between variables may not be clear. This is not the ideal approach for a process such as linguistic analysis, which depends on rigorousness and certainty in our conclusions so that progress can be made with the data extracted.

There are two main approaches which define research methods. These are qualitative methods, typically associated with interpretivism and quantitative methods, associated with positivism. Quantitative research focuses specifically on numerical data which measure the scale, range or frequency of phenomena. The data will be analysed and statistically treated to see whether to reject a hypothesis or not. On the other hand, qualitative research examines aspects of the subject which include values, attitudes and perceptions. Thus, neither is weaker or stronger than the other, merely suited to different types of research. Both qualitative and quantitative methods are designed to address a specified research question, but both contribute to a different aspect. The qualitative method will allow the researcher to investigate and gather a better understanding of complex phenomena, whereas the quantitative method will provide an objective measure of reality. In the mixed methods approach, which uses both qualitative and quantitative methods, researchers collect numerical data and analyse that, but will only be able to answer the research questions with the additional help of the collection and examination of narrative data. This provides answers about the complicated nature of phenomenon from the participants' perspective, as well as the relationship between measurable variables (Brannen, 2005).

- Research strategy

The research strategy adopted to address the topic was an experimental one. This, as noted by Saunders *et al.* (2009), is a strategy that is also closely connected to the natural sciences and traditional scientific research. Thus, the choice of experimental research lends itself well to a positivist philosophy and a deductive approach. The experimental strategy also lends itself well to research where the data is sensitive to errors, as it allows the researcher to account for possible variables and ensure that no errors are made. An experiment describes any situation where researchers establish groups of categories for the subjects being observed and tests them, so as to compare their results. In our case we are comparing the results of our system's data retrieval to the results of prior systems.

1.6 Research Design

Due to the nature of the data and the use of the probabilistic method (BBN), this research adapts a quantitative approach in the development, implementation, evaluation, and validation stages. Bryman (2004) outlines several steps to quantitative research (Figure 1.2) Cohesiveness and thoroughness are essential for a study using this approach. Bryman notes that it is rare that the process is applied exactly as described. Due to difficulties in collecting and processing data at every step, most research will meet hurdles that alter the path. However, the goal of the researcher is approximate this approach as closely as possible.

Step one, beginning with theory, indicates that there is a strong deductive connection between theories and research, that is that the researcher will begin with a concept, which can then be formed into a hypothesis at step two. Only once the hypothesis is outlined can the researcher begin to conduct his/her research. However, Bryman notes that often a theory without a hypothesis can be enough for research to begin. For our research, which is primarily experimental, a hypothesis is substituted by research questions in order to guide the experiment stage. The third step, research design, presents a hurdle to researchers, as Bryman notes, and must be undertaken carefully.

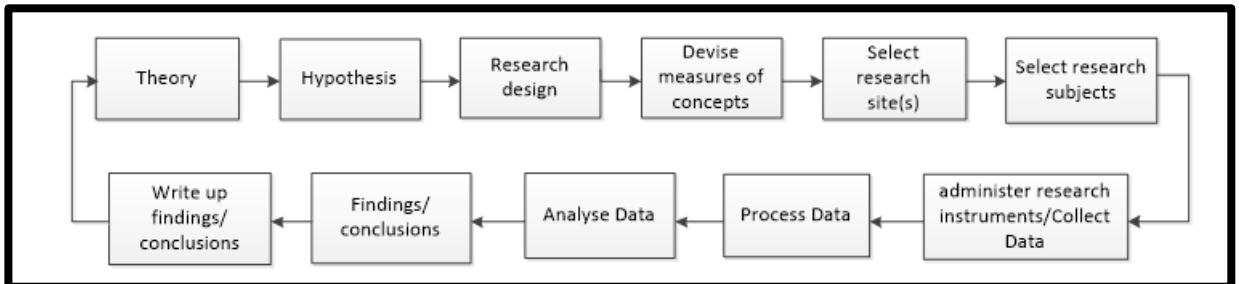


Figure 1.2 The main steps in quantitative research (Bryman, 2004).

The type of design chosen may affect the validity of findings or of analysis and conclusions drawn. At step four the measures of concepts are devised. This involves taking real world concepts, transforming them into their measurable elements, and defining the parameters of the research being conducted. Step five is the selection of research sites, which may be followed or combined with step six, the selection of subjects. Proper selection of subjects and a suitable platform are essential to survey and interview research; however, in this research the proper selection of corpus is important. Step seven follows with the administration of research instruments and the collection of data. For our research this entailed the preparation of the tools to be used and the input of the corpus into the tools, as well as any necessary pre-processing. Step eight covers the processing of data, where the results are collected and sorted.

Step nine addresses analysis. This step is highly important and must be conducted effectively. The researcher narrows down the data into either representative samples or statistics, giving a smaller number from which to test the relationships between variables and extrapolate meaning. Despite validation not being a part of the diagram as outlined by Bryman, a k-fold cross validation is also conducted in our research to validate the system. The researcher's interpretation of the results then goes into developing a conclusion to present alongside the raw findings. Finally, at steps ten and eleven the findings and conclusion are determined and written. Figure 1.3 illustrates the main steps in our research.

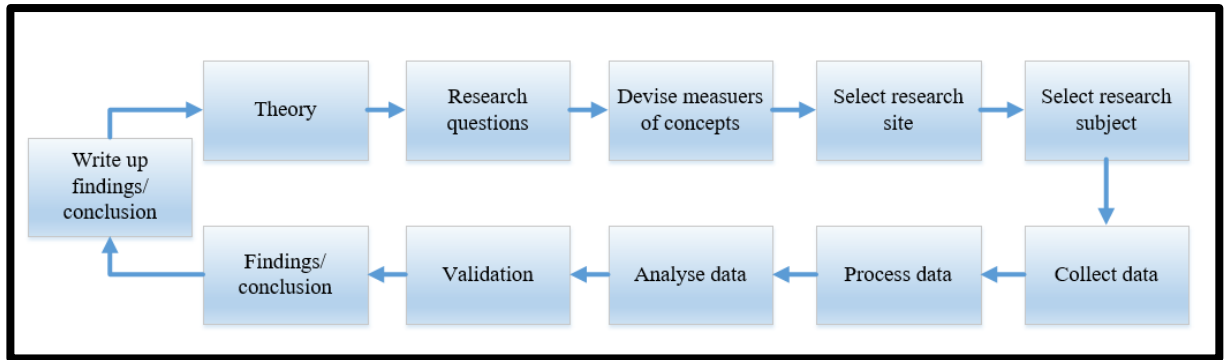


Figure 1.3 The main steps in our research

1.7 Challenges of Arabic Language Processing

The majority of research efforts on NER have been devoted to English language texts while fewer research projects have investigated the field of Arabic text. Since the Arabic language is the mother tongue of more than 300 million citizens in more than 25 countries, devoting more research to undertake NER for Arabic texts is crucial (Shaalán, 2010). Arabic has many traits which make building an effective NER system a very challenging task. Some of these challenges pertaining to our study are described below:

- 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 (Farghaly and Shaalan, 2009).

- The ambiguity of the Arabic language

According to (Attia, 2008), there are many levels of ambiguity in Arabic language which makes the computational process a challenging task. Farghaly and Shaalan (2009) highlighted six levels of ambiguity in Arabic language including: homographs, internal word structure ambiguity, syntactic ambiguity, semantic ambiguity, constituent boundary ambiguity, and anaphoric ambiguity. For our system, the homographs levels of ambiguity can be considered one of most challenging obstacles. For instance, the word “سرطان” could mean the disease *cancer* or the animal *crab*, the word “ألم” could mean *pain* or the question *haven't you*, and the word “نقص” could be a noun that means *loss* in the phrase *weight loss*, could be a verb that means *tell a story*, or a verb that means *cut*.

Some of the challenges which are related to general Arabic NER tasks are:

- 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 often written without diacritics (Alkharashi, 2009).

- Lack of Capitalisation

Languages such as English use capitalisation and most named entities begin with a capital letter. However, in the Arabic language capitalisation does not exist as an orthographic feature in relation to identifying named entities which are proper names, acronyms and abbreviations (Farber et al., 2008). 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 (Algahtani, 2011).

1.8 Ethical Issues

This research project was conducted in full compliance with the ethical regulations of Staffordshire University. The University's Ethical Review Policy has been consulted and that all ethical issues and implications in relation to this research project have been considered.

1.9 Thesis Structure

The rest of this thesis is structured as follows:

Chapter 2 surveys the current research in the field of Arabic named entity recognition including the NER approaches, the evaluation metrics, and the NER tools and resources for Arabic language. Chapter 3 discusses the theoretical foundations that underpin our research which are the natural language levels and Bayesian belief networks. Chapter 4 focuses on the first stage of our NER system which is the natural language processing stage. Chapter 5 describes the second stage of our system which include the application of BBN approach to the classification and recognition of appropriate named entities. Chapter 6 concludes the research developed in this thesis and highlights the research findings and its original contributions to the knowledge.

Chapter 2: Literature Survey

2.1 Introduction

Research into Arabic named entity (NE), until recently, had not been investigated in much depth. In part, this was due to the lack of available digital resources, and in part, it was due to the complicated structure of the Arabic language. However, in the last decade, Arabic digital texts have expanded in number, encouraging the developers and researchers of named entity recognition (NER) systems to make use of the Arabic language to develop better natural language processing methods. The focus of this chapter is primarily related to the research into Arabic NER system.

2.2 Evaluation Metrics of NER

As NER systems are still developing, it is important to test the accuracy of every system in every domain. There are several ways of evaluating NER, and all of them use some point of comparison between an established goal accuracy of 100% based on manually annotated data and the accuracy of the method being tested. To assess NER system accuracy, three evaluation metrics are normally used. Precision, recall, and F-measure are used for the most reliable, academically sound results (Chinchor *et al.*, 1998; Marsh and Perzanowski, 1998). These measures rely on comparing and contrasting the extracted entities with the manually annotated corpus. Precision compares the correct entities that have been recognised (true positives) to the total number of recognised entities (true positives and false positives) to give a percentage of accurate identification. While recall compares the correct entities that have been recognised (true positives) to the total number of correct entities (true positives and false negatives), F-measure is a harmonic mean that gives equal weight for recall and precision.

$$Precision = \frac{True\ Positives}{True\ positives + False\ Positives} \quad \text{Equation 3.1}$$

$$Recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad \text{Equation 3.2}$$

$$F\text{-measure} = \frac{2 * recall * precision}{recall + precision} \quad \text{Equation 3.3}$$

Figure 2.2 provides a visualisation aid to illustrate how precision and recall are measured using true positive, false positive, and false negative notations. While this is a fairly standard way of assessing accuracy and one of the cheapest and most efficient and therefore is one of the best evaluation metrics used, it is important to remember that there are three different standards involved in performing the previous evaluation metrics when there are boundary errors in entities represented by more than one word. These standards are MUC, CoNLL, and ACE. The MUC (Chinchor and Robinson, 1997) uses tolerant standards in which the NER system is receiving partial credit when a partial match with the NE occurs. In contrast, CoNLL (Tjong Kim Sang and De Meulder, 2003) uses strict standards where only an exact match will be credited. In addition, ACE (2004) standards are even more strict as they consider other parameters like co-reference resolution and mention detection.

2.3 Applications of NER

Named entity recognition can be used to support research in different natural language processing systems, namely text clustering, question answering, information retrieval, text summarisation, machine translation and especially for information extraction purposes. For instance, NER was used to develop a question answering system (Lee *et al.*, 2006) to improve machine translation quality (Babych and Hartley, 2003; Aulakh and Kaur, 2014), to develop a sentiment analyser (Nakov *et al.*, 2016), and to develop a topic detection system (Menner *et al.*, 2016). In languages besides Arabic, NER has abundant applications in various other research fields, for instance, in molecular biology (Krallinger and Valencia, 2005), bioinformatics (Leaman and Gonzalez, 2008), and medicine (Jimeno *et al.*, 2008).

The NER community has developed invaluable tools and systems to be applied across various fields such as the Stanford named entity recogniser, as well as to support natural language processing research where NER has become a cornerstone of this field. Of course, these systems have their own unique problems. Even the most modern NER systems can be fragile, as they have limited transferability. Due to this, NER systems tend to be highly domain specific (Al-Shalabi *et al.*, 2009), meaning that they cannot be then transferred to another domain except in the most rudimentary manner.

The vast majority of Arabic NER work focuses on newswire texts for different domains, mainly because of availability and digital access for further processing. One of the earliest Arabic NER research was carried out by Maloney and Niv (1998) who used newspaper articles in Arabic to test their TAGARAB Arabic name recogniser to isolate person names, locations, dates, times, and numeric entities from surrounding words. In 2009, Traboulsi investigated Arabic NER in financial news text

and extracted person names. Aljazera website articles were used in Elsebai's study (2009) to extract person names, organisation, location, time, date, and monetary entities from the political domain. The *Al-Raya* newspaper also was used by Al-Shalabi *et al.* (2009) to extract location, person name, temporal event, equipment, and scientific organisation, while *Assabah* and *Alanwar* newspapers were used by Zaghouni *et al.* (2010) to extract person names, location, date, number, and quotations. The newswire texts were used also by Alruily (2012), Asharef and Omar (2012), and Al-Shoukry and Omar (2015) to apply NER to extract entities related to the crime domain such as person names, location, organisation, date, and time.

In addition, many Arabic corpora that are newswire text-based were used extensively by different researchers to implement and test their NER systems. For instance, ANERcorp, which comprises 136 newspaper articles, is applied in studies conducted by Benajiba *et al.* (2007), Zaghouni (2012), Al-Jumaily *et al.* (2012), and Shihadeh and Neumann (2012). ANERcorp was usually used to extract person names, location, organisation, and miscellaneous entities. The ACE (2003, 2004, 2005) corpora, which are composed of newswire and broadcast news data, were extensively applied in studies conducted by Abdul-Hamid and Darwish (2010), Benajiba *et al.* (2010), and Oudah and Shaalan (2013). ACE corpora were usually used to recognise person names, location, organisation, vehicle, and weapon. Details of ANERcorp, ACE corpora, and other Arabic corpora are provided in Section 2.6.1.

Despite the focus on newswire text, other researchers analysed texts from different resources. For instance, texts from social media were used by Omnia and El-Beltagy (2012) and Zirikly and Diab (2015) in which they expanded their research to cover colloquial and dialectical Arabic. Texts from religious books were used by Bidhendi *et al.* (2012) and Alhawarat (2015). However, although research in Arabic is scarce in all domains except the newswire domain, NER in the domain of medical texts in Arabic is even less explored. The medical domain was examined by Samy *et al.* (2012), who establish two strategies in order to extract medical terms. However, no further details have been provided about these terms. Moreover, the evaluation of the strategies used a small dataset containing only 2,273 tokens.

The number of available NER systems, paired with the current lack of research on Arabic language texts and the need for further development of word lists and systems, makes this an important field for further research. That said, research into NER has been applied to medical texts in many other languages. Medical and biomedical NER in English and Chinese has made fast progress, especially in more recent years, and is becoming invaluable in many areas of medical research, enabling information extraction and knowledge discovery from vast corpora.

Bodenreider and Zweigenbaum (2000) utilised ‘fine grained’ NE sub-categories applied to English medical texts, extracting information about medication names and disease names. Rindfleisch *et al.* (2000) observed the extraction of drug NEs. In 2002, a corpus, GENIA, was developed, categorising the NE types of proteins, DNA, RNA, cell lines, and cell types, resulting in the later development of further corpora, such as Tsuruoka and Tsujii’s 2003 categorisation of proteins and categorisation of chemical names from the same year by Narayanaswamy *et al.* (2003).

In 2008, Leaman and Gonzalez (2008) went on to conduct a survey on the advances in biomedical NER. Their open-source, executable survey of the advances, which is called BANNER, is presented to serve as a benchmark for the other NER systems. BANNER displayed high performance by employing a combination of the most established and most recent techniques in the field for the time.

In 2010, Sondhi also conducted a survey of available NE extraction techniques in the biomedical domain, finding that that “the use of a standard set of features along with machine learning techniques may no longer be enough to improve performance” (Sondhi, 2010, p. 10). Contextual features were found to further boost performance. Sondhi also noted that the GENIA corpus and its inconsistencies as well as the need for exact matching criteria unnecessarily reduced performance ratings, demonstrating the need for the development of an application-specific evaluation as well as newer stronger corpora. Beyond English, Chinese texts have also made use of adapted versions of the same systems, employing a deep neural network for the analysis of electronic health records (Wu *et al.*, 2015).

In 2013, Bodnari *et al.*, working from MIT, presented their participation in Task of the CLEF eHealth challenge. They tested their NER system, using conditional random fields (CRFs). Their goal was to identify disorder NEs from available electronic medical records. They found that “a rich feature set and external knowledge gathered from specialized terminologies and general domain knowledge repositories” (Bodnari *et al.*, 2013, p. 7) gave good precision results but poor recall results. The overall all F-measure was 59.8% in the context of strict evaluation and 71.1% in the context of relaxed evaluation.

Zhang and Elhadad (2013) were the pioneers of unsupervised biomedical NER in 2013. Testing their system on the i2b2 and GENIA corpora, two accepted datasets, they utilised seed knowledge to improve classifications based on what they called ‘signature’ similarity. Their results held true to previous analyses performed on the same corpora, proving that their system was effective and easily generalised to other English medical texts.

In 2015, Lee *et al.* (2015) assessed a Conditional Random Fields system for its effectiveness recognising Disease Named Entities across a corpus constructed from English PubMed articles. They

achieved an accuracy of 86.64% for their F-measure, thanks to a very high precision rating. They concluded that the system had benefited from the adjustments they used to optimize it, and that further normalization along the same lines may result in even greater improvements.

In 2016, Wang *et al.* also observed the effectiveness of Conditional Random Fields combined with a Support Vector Machine for the task of NER in Chinese medical texts. The results were strong, with F-measures ranging from 78.47% to 94.58%, with longer named entities such as Treatments being less accurate. They concluded that machine learning was a strong and valuable tool when it came to extracting Named Entities from digital medical records in Chinese.

To the best of our knowledge the research into NER focusing on the Arabic language is limited to the domains politics, economy, and crime, and the texts used were mainly based on newswire resources. Furthermore, the analysis of Arabic medical documents has not been explored as yet. This has provided yet another motivation to test our approach to this unexplored domain. Moreover, this work supports and highlights the lack of similar work in the field of Arabic medical NER.

2.4 NER Approaches

There are three main approaches adopted by NER researchers. They are divided into the most labour intensive approach, which is the rule-based hand-written approach, the machine learning (ML) approach, and a combination of the two (Figure 2.1). Table 2.1 summarises the main findings which are discussed in the following sections.

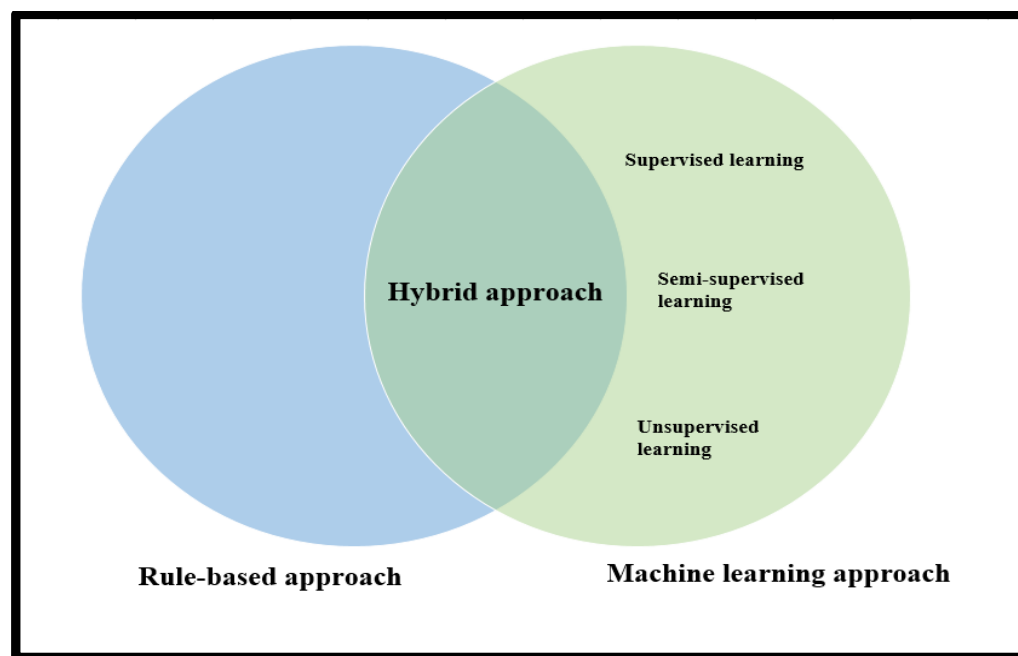


Figure 2.1 NER approaches.

Table 2.1 A Summary of the related Arabic NER work.

Publication	Method	Corpus	Entities	Results
(Maloney and Niv, 1998)	Rule-based	Own corpus (14 texts from the Al Hayat news paper)	Number, entity, time, Location, and Person	90.9% F measure For training set 85% F measure for blind set
(Shaalán and Raza, 2007)	Rule-based	ACE and Treebank Arabic data sets	Person names	87.5% F measure.
(Al-Shalabi et al., 2009)	Rule-based	Own corpus (20 articles from Al-Raya newspapers)	Location, person names, event, organisatoin, temporal, equipment, and scientific	86.1% Precision
(Elsebai, 2009)	Rule-based	Own corpus (one thousand articles from Aljazeera website)	Person name, organisation, location, time, date, and money	Results in F-measure Person name (89.67%) Organisation (86.52%) Location (85.87%) Time (96.14%)

Publication	Method	Corpus	Entities	Results
				Date (93.82%) Money (94.59%)
(Zaghouani et al., 2010)	Rule-based	Own corpus (35 news articles from the newspapers Assabah and Alanwar.	Person, organisation, location, date and time, and numeric expression	74.95% F-measure
(Elsebai and Meziane, 2011)	Rule-Based	(500 articles from Aljazerra website)	Person names	89% F-measure
(Zaghouani, 2012)	Rule-Based	ANERcorp	Person, organisation, and location	67.13% F-measure
(Al-Jumaily et al. 2012)	Rule-Based	ANERcorp	Person, organisation, and location	Person 77.27% Location 70.87% Organisation 57.30%
(Asharef et al., 2012)	Rule-Based	95 Arabic crimes articles from four Arabic newspapers	person names, locations, organizations, dates and times	89.46% F-measure

Publication	Method	Corpus	Entities	Results
		(Albyan, Aljazeera, Okad and Gorena)		
(Alruily, 2012)	Rule-Based	Arabic Crime News Report Corpus (ACNRC)	Crime type, location (scene) and nationality	Crime type (69%) Location (95%) Nationality (88%)
(Shihadeh and Neumann, 2012)	Rule-Based	ANERcorp	Person, location and organisation	30% F-measure
(Benajiba and Paulo, 2007)	Maximum entropy (Supervised learning)	ANERcorp	Person, location, organisation, and miscellaneous	54.11% F-measure (without using ANERgazet) 55.23% F-measure (using ANERgazet)
(Benajiba and Paulo, 2008)	Conditional Random Fields (Supervised learning)	ANERcorp	Person, location, organisation, and Miscellaneous	65.91% F-measure
(Benajiba et al., 2008)	Support Vector Machines	UPVCorpus ACE (2003 -	Person, location, organisation, and	82.71% F-measure

Publication	Method	Corpus	Entities	Results
	(Supervised learning)	2004-2005)	miscellaneous	
(Abdul-Hamid and Darwish, 2010)	Conditional Random Fields (Supervised learning)	ANERcorp ACE (2005)	Person, location, and organisation	81% F-measure (ANERcorp) 76% F-measure (ACE)
(Koulali and Abdelouafi, 2012)	Support Vector Machines (Supervised learning)	ANERcorp	Person, location, organisation, and miscellaneous	83.20% F-measure
(Mohammed and Omar, 2012)	Artificial Neural Network (Supervised learning)	ANERcorp	Person, location, organisation, and miscellaneous	92% Accuracy
(Bidhendi et al, 2012)	Conditional Random Fields (Supervised learning)	Own corpus (three ancient Arabic religious and historical books: Seffeyn, Al-Irshad, and Sharaye)	Person, location, organisation, and miscellaneous	99.93 F-measure (Seffeyn) 93.86 F-measure (Al-Irshad) 75.68% F-measure (Sharaye)
(Morsi and Rafea, 2013)	Conditional Random Fields (Supervised learning)	ANERcorp	Person, location, organisation, and	68.05 F-Measure

Publication	Method	Corpus	Entities	Results
	learning)		miscellaneous	
(Shabat and Omar, 2015)	Naïve Bayes, Support Vector Machine and K-Nearest Neighbor classifiers(Supervised learning)	Own corpus (extracted from the Malaysian National News Agency)	Weapons, nationality, location, and type of crime	89.48% F-measure (crime type) 93.36% F-measure (crime-related entities)
(Alotaibi, 2015)	Maximum entropy, and Conditional Random Fields (Supervised learning)	NewsFANEGold, WikiFANEGold, and WikiFANEAuto	8 coarse-grained classes (Person, organisation, location, geo-political, facility, vehicle, weapon, and product) and 50 fine-grained classes	61.67 % F-measure (NewsFANEGold) 54.04 % F-measure (WikiFANEAuto) 71.84% F-measure (WikiFANEAuto)
(AbdelRahman et al.,2010)	Conditional Random Fields (Supervised learning) and bootstrapping (Semi-supervised learning)	ANERcorp	Person, location, organisation, job, device, car, cell phone, currency, date, and time.	Person (67.80%), location (87.80%), organization (70.34%), job (69.47%), device (77.52%), car (80.95%), cell phone (80.63%), currency (98.52%), date (76.99%), and time (96.05%).
Althobaiti et al. (2013)	Bootstrapping (Semi-supervised learning)	ANERcorp and ACE (2005)	Person, location, and organisation.	Person (64.14%), location (73.06%), and

Publication	Method	Corpus	Entities	Results
				organization (54.52%).
Darwish and Gao (2014)	semi-supervised twopass method	ANERcorp and own corpus (1,423 tweets)	Person, location, and organisation.	65.2% F-measure
Althobaiti (2016)	Semi-supervised method and distant learning method	ANERcorp	Person, location, and organisation.	Semi-supervised method (64.27%) distant learning method (64.92%)
Abdallah et al. (2012)	Decision trees and rule-based method (Hybrid approach)	ANERcorp and ACE (2003)	Person, location, and organisation	88.87% F-measure
Meselhi et al.(2014)	Support Vector Machine and rule-based method (Hybrid approach)	ANERcorp	Person, location, and organisation.	Person (96.65%) location (94.8%) organisation (92.9%)

2.4.1 Rule-based approach (the hand-written approach)

The vast majority of early NER research was performed using the rule-based approach. This was due to the inadequacy of the systems that existed at the time. However, as ML systems have improved, the amount of research conducted exclusively using a rule-based approach has dwindled. The rule-based method depends on hand-made linguistic rules (such as grammar), which are defined by linguists. It has been used extensively in many studies (e.g., Maloney and Niv, 1998; Shaalan and Raza, 2007; Al-Shalabi *et al.*, 2009; Elsebai *et al.*, 2009; Elsebai, 2009; Zaghouani *et al.*, 2010; Zaghouani, 2012; Al-Jumaily *et al.*, 2012; Elsebai and Meziane, 2011; Asharef *et al.*, 2012; Omnia and El-Beltagy, 2012; Alruily, 2012; Shihadh and Neumann, 2012).

One of the earliest Arabic NER research studies is Maloney and Niv's 1998 study. They introduced the TAGARAB system, which uses a morphological analyser in order to decide where a non-name context starts and where a name ends. An evaluation was conducted where 14 texts were picked randomly from the Al Hayat CD-ROM and manually tagged. The outcome of this was 89.5% precision, 80.8% recall, and 85% F-measure. Details of those well-known measurements are presented in Section 2.5.

In 2007, Shaalan and Raza (2007) developed PERA which is a NER system that identifies person names. The PERA system includes three components. First, a gazetteer provides a whitelist of person names that, regardless of the grammar, will extract the exact matching NEs. Second the text is analysed using grammar rules. Finally, there is the filtering mechanism that is used to exclude invalid person names from the NEs. The ACE and Treebank Arabic data sets were used to evaluate PERA. It achieved an 85.5% for precision, 89% for recall, and 87.5% F-measure.

In 2009, Al-Shalabi *et al.* (2009) noted once more that proper names present a challenge to information retrieval, particularly cross language information retrieval, and even more so when it comes to extracting proper nouns in Arabic. They noted the importance of context and domain in the task of information retrieval in Arabic. They also added that not only do the lack of capitalisation and the use of common nouns and adjectives in names confuse matters but also many Arabic names are presented in a long series of names, all of which require appropriate categorisation. Al-Shalabi *et al.* (2009) evaluated their system using 20 articles from *Al-Raya* newspapers, and a precision rating between 75% and 91.6% was achieved. However, the limited applications and labour intensity of this process suggest that it may not be the most effective method for expanding into other domains.

In 2009, Elsebai (2009) suggested a tool for natural language processing in modern standard Arabic that would take advantage of the recent abundance of electronic documents available in Arabic. Focusing on NER, Elsebai suggested a rule-based system to create “an efficient and effective

framework for extracting Arabic NEs from text” (Elsebai, 2009, p. 11). The approach was designed to make use of contextual and morphological information to identify NEs. The context was assessed based on words that represented clues for each observed NE type. Morphological information identified the part of speech for each word. Elsebai developed and implemented rules to recognise the position of the NE. He based the system architecture on the GATE system, as it was considered the most robust as well as being available free of cost. Consequently, Elsebai built the Buckwalter Arabic Morphological Analyser (BAMA) over GATE. Manual corrections were required to create more accurate results. Trigger word lists were also necessary for the identification of proper nouns. Faced with a problem that many other researchers have faced in the field of Arabic NER, he had to build a corpus, as free annotated Arabic corpora could not be located for the task. He chose a selection of texts from the *Aljazeera* website for the corpora, up to a thousand articles. The achieved results varied between 85.87% and 96.14% F-measure depending on the type of NE.

Zaghouani *et al.* (2010) investigated the use of a low-resource language-independent NER system named Europe Media Monitor (EMM)-NewsExplorer with Arabic text. They noted that many search systems use such programs for the analysis of news texts; however, these programs are usually either monolingual or less accurate. They suggested that alterations would need to be made to make these systems properly adapted to the Arabic language. They found that, under the existing systems, often precision had to be optimised, but recall was sacrificed in the effort. However, they expressed a wish to improve the recall for future work. Later, in 2012, Zaghouani (2012) again investigated the use of EMM-NewsExplorer for the extraction of person, organisation, and location NEs from news web sources, such as newswire. The overall result is 73.39% for precision, 62.13% for recall, and 67.13% for the F-measure.

Elsebai and Meziane (2011) applied a rule-based approach to extract person NEs from news articles located on the *Aljazeera* website. They ran the system twice, first with 700 news articles from *Aljazeera* and then with another 500 news articles from *Aljazeera*. They used keywords instead of complex grammar or ML techniques. Despite the system being so simple, they achieved a good result with a total precision of 93%, recall of 86%, and F-measure of 89% for the first trial and 88%, 90%, and 89%, respectively, for the second attempt.

A more recent rule-based NER system was developed by Al-Jumaily *et al.* (2012). This system used GATE along with different gazetteers from GATE, DBPedia32, and ANERGzet.33, while ANERcorp was used to evaluate the system. Two experiments were conducted to determine the effect of Arabic prefixes and suffixes in terms of recognition. If Arabic tokens (prefix-, stem, and -suffix) were recognised, tests were then done to determine how compatible they are between the three (prefix-stem, stem-suffix, and prefix-suffix). The outcome was that the verification process had improved all of the

results of all NE types, but none of them in a symmetrical manner. Precision had gone up by 7.32% for person, by 5.55% for location, and by 5.14% for organisation.

Asharef *et al.* (2012) applied NER to the identification of NEs in crime documents, a specialised area. By employing a rule-based approach that utilised morphological information and predefined crime and general indicator lists in a NER system, they devised appropriate rules and patterns for the domain. The system's accuracy was 90%, indicating that the system was effective, even in an understudied domain.

In 2012, Alruily explored the application of text mining techniques in Arabic text. Alruily noted that, until that point, such techniques had only been applied to English and generally to specific domains. Alruily indicated that there were few mining techniques being used in Arabic and that most domains were completely unaddressed. Alruily's system, the Crime Profiling System, would assist in identifying important information and extracting it. Due to the constraints of the field, Alruily conducted the research without predefined dictionaries or an annotated corpus. This resulted in a self-organising map approach, which would perform the clustering based on rules written concerning crime type, location, and nationality (Alruily, 2012).

In 2012, Shihadeh and Neumann (2012) presented ARNE, a pipeline software for the task of Arabic NER. It "includes tokenization, morphological analysis, Buckwalter transliteration, part of speech tagging and named entity recognition of person, location and organisation named entities" (Shihadeh and Neumann, 2012, p. 24). Using a simple, fast, language-independent gazetteer lookup method, they used the morphological analysis from the pipeline to remove affixes and improve performance. However, the results were fairly discouraging with 38% for precision, 26% for recall, and 30% for F-measure. They recognised the need for a part-of-speech tagger, which may bring their accuracy up to modern standard rates.

2.4.2 Machine-learning approach

Machine learning is another approach extensively used to develop statistical models for NE prediction. Although, at first, the accuracy of ML made it unreliable, the accuracy levels now match and even outstrip those of rule-based approaches, all while reducing the cost of conducting the research in terms of time and money. The ability to recognise unknown NEs for what they are and to classify them automatically allows much more research to be conducted. The use of training examples as opposed to hand-written rules has opened many doors, as evidenced by the fact that, in the Message Understanding Conference 7 (MUC 7) competition, five systems out of eight were rule-based systems, whereas, at the Conference on Computational Natural Language Learning 2003 (CoNLL 2003), 16 systems were presented. However, it is important to note that where ML fails to deliver the correct

results, many researchers will fall back on writing rules to adjust the system. The ML method can be split into three categories: supervised, semi-supervised, and unsupervised learning.

2.4.2.1 Supervised learning

Supervised learning (SL) includes studying and analysing both positive and negative features of NE examples from a broad collection of annotated corpora. Techniques that belong to SL are maximum entropy (ME) models, Conditional Random Fields CRFs, support vector machines (SVM), and neural networks, which are applied equally to Arabic texts. Maximum entropy was applied by Benajiba *et al.* (2007), while CRFs were applied by Benajiba and Rosso (2008), Abdul-Hamid and Darwish (2010), Bidhendi *et al.* (2012), Morsi and Rafea (2013), and Alotaibi (2015). Benajiba *et al.* (2008) and Koulali and Abdelouafi (2012) implemented SVMs. Benajiba *et al.* (2009) 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. They also explained that the choice of the appropriate features is a very significant phase of any ML-based system. Mohammed and Omar (2012) adapted neural networks in their approach, making use of the back propagation training algorithm.

In 2007, Benajiba and Paulo (2007) explored the results of their ANERsys, an Arabic NER system, to improve the precision results that it had achieved in its initial version. Their system was initially built based entirely on a ME approach and trained on their own annotated corpora. The results were in some ways encouraging, as they found ME was fairly successful at categorising NEs in Arabic texts.

In 2008, Benajiba and Paulo (2008) continued to explore the uses of their ANERsys. Its performance had not initially scored as highly as they would have liked, and in this study, they set out to raise its performance. This time, they proposed that they would improve the accuracy of ANERsys by swapping from a ME probabilistic model to a CRF model. Their first version of ANERsys, ANERsys 1.0, involved ME. It had already been successful at classification tasks; however, error analysis showed that the system, like many others, could not adequately handle NEs with multiple tokens, making it challenging to successfully recognise NEs which consist of more than one token. In this second version of the system, ANERsys 2.0, they swapped to a two-step approach. The first step would be the detection of the start and closing tokens of the NE, and the second step would be classification. Tokenisation of the data improved the results. The addition of four different combined gazetteers further improved the results. The features used, however, were not specific to the Arabic language but were language-independent. These preliminary experiments showed a 10-point improvement from ANERsys 1.0 to ANERsys 2.0, from no features to all features.

Taking advantage of the developing resources for Arabic NER, in 2008, Benajiba and Paulo, now associated with Diab, explored the uses of SVMs. They made use of a ML framework and a

combination of language-independent and language-specific sets of features and found that combining all possible features yielded the best results so far for ANERsys (Benajiba *et al.*, 2008).

In 2009, Benajiba *et al.* (2009) later investigated the effect of different feature sets in the three ML systems based on ANERsys: SVMs, ME, and CRFs, with a focus on Arabic NER in the domain of broadcast news data. They found that a combination of 15 features resulted in the highest accuracy in terms of performance, suggesting again that the addition of more features disproportionately benefits Arabic NER.

In 2010, Abdul-Hamid and Darwish (2010) tested a simplified feature set to assist in accurately identifying NEs in Arabic texts without using morphological or syntactic analysis or gazetteers. They employed a CRF sequence labelling model trained on specific features. They compared the results to others in previous studies where Arabic specific features, such as part-of-speech tags, were employed and found some improvements and some loss of F-measure accuracy depending on the category of the observed NE. Their goals were to identify simplified features to make the task of Arabic NER easier, to use leading and trailing character n-grams in words, which would capture valuable clues indicating the presence of NEs, and to incorporate word language modelling features to capture NE word association and distribution. Their sets were effective despite the simplicity and overcame some of the issues caused by the complexity of the Arabic language. The results were as accurate overall as previous methods, albeit with a bias against locations and in favour of organisations and people. These results suggest that simpler methods of conducting Arabic NER may be possible.

In 2012, Koulali and Abdelouafi (2012) have developed a NER system called ANER which is based on SVM with a set of features. They evaluated the effect of the combination of these features on the performance of ANER. Automatic extraction of patterns was used to enhance the performance of their system. The ANER system has achieved an average of 83% F-measure. In 2012, Mohammed and Omar proposed a novel solution in the field of Arabic NER: the application of an artificial neural network. The main task of a neural network approach is to learn to recognise the component patterns of a text automatically, enabling intelligent decision making based on the available data. This approach is a ML approach to the classification of Arabic NEs. The approach was divided into three phases. The first phase addressed the accuracy in precision, recall, and F-measure for each class of NE addressed by the artificial neural network. Each class showed some differences in the accuracy of each measure, with precision being the strongest. This first phase also addressed decision tree accuracy. Although many measures were stronger than the artificial neural networks, the overall results were less consistent, and the total accuracy of the artificial neural network was higher across all text volumes. The branching of artificial neural networks into this domain achieved a 92% accuracy result. Compared to a decision tree system using the same data, the artificial neural network outperformed the decision

tree, which only achieved 87% accuracy. Bidhendi *et al.* (2012) applied a supervised ML approach to extract person NEs from ancient Islamic texts. They made use of CRFs and proposed the system they called ‘Proper Name candidate injection’ as part of the process. The declared results from this method are very high for historical and traditional data with 99.93% and 93.86% F-measure respectively.

In 2013, Morsi and Rafea (2013) used a supervised ML approach to assess the effect of different features on the performance of Arabic NER conducted via CRF models. Morsi created a baseline and the best result from the various feature combinations was over 10 points above the baseline with a 68.05% F-measure.

In 2015, Shabat and Omar (2015) applied naive Bayes, SVM, and K-nearest neighbour classifiers as base classifiers in the task of NER in the domain of crime news. They extracted NEs classified as crime type or crime-related. Then, they applied a weighted voting ensemble method to combine the results of all three classifiers. Their final results are tested against manually annotated data from BERNAMA. The end result is an F-measure of 89.48% for identifying crime type NEs and 93.36% for identifying crime-related NEs. Alotaibi (2015) applied supervised ML to the extraction of NEs in fine-grained classes, as opposed to coarse-grained classes. Maximum entropy and CRFs were applied to extract NEs from 50 sub-classes. A corpus was built based on Arabic language *Wikipedia* articles. Two highly rated corpora from other domains were also used in the evaluation process. The results are three fine-grained corpora and a fine-grained gazetteer specific to Arabic *Wikipedia* (Alotaibi, 2015).

Roth and Yih (2002) developed a method for recognising relations and named entities in English texts, accounting for mutual dependencies. They compared three different classifiers: basic, omniscient, and BN. These classifiers are learned independently using local features and are able to predict entities and relations separately. Their approach is to extract two named entities: person, and location and two relations: kill and born-in. They found that their belief network approach decreased recall but significantly improved precision. They claim that knowing the class labels of relations has not improved significantly the entity classifier mainly as the difference of Basic and Omniscient approaches is not very significant and usually less than 3% in terms of F1 in all three datasets.

Jochin *et al.* (2014) addressed the extraction of risk events and probabilities from biomedical texts. They investigated the determination of the parameters of a BBN using conditional random fields (CRFs) and were the first to research this matter as a sequence tagging problem, labelling spans of text as events. For this purpose, their corpus consisted of 200 free abstracts extracted from PubMed. They observe that risk events are fairly heterogeneous and have greater semantic variety than bio-molecular entities, which made them stand out. They note it is difficult to extract conditional probability

statements, due to their variety of forms. Their CRF approach improves over the established baseline, proving such a task is best handled as a sequence tagging problem.

2.4.2.2 Semi-supervised learning

Semi-supervised learning (SSL) was a recent development in 2007, and almost 10 years later, it remains far less explored than SL. ‘Bootstrapping’ is considered the main technique of SSL. It makes use of ‘seed’ examples of a NE category, where the context in which the example is used is broadened and employed to identify other NEs in the same category. This simplifies the process of locating and categorising new NEs in a large corpus. 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 that have similar context. The learning process can then be reapplied to the sentences that have been found so that it can come up with new and relevant context. With the repetition of this process, the system will recognise a broad number of disease names as well as context (Nadeau and Sekine, 2007).

Semi-supervised learning is a less explored field in natural language processing. Therefore, there has also been very limited application of it to Arabic natural language processing. The most common method, as elsewhere, is bootstrapping. AbdelRahman *et al.* (2010) combined CRF with bootstrapping to extract a wide range of entities that include person, location, organisation, job, device, car, cell phone, currency, date, and time. The ANERcorp dataset was also used to evaluate the system for all entities except device, car, and cell phone. The results show that the F-measure varies between 69.47% and 96.05%, depending on the type of entities. Althobaiti *et al.* (2013) adapted the bootstrapping algorithm in their system (ASemiNER) in order to identify specific entities, such as person, location, and organisation. The system can also recognise specialised entities, such as politician names, sport persons, and artists. The outcome of the F-measure was 64.14%, 73.06%, and 54.52% for person, location, and organisation, respectively.

In 2010, Rahman *et al.* (2010) analysed various integrated ML techniques when applied to Arabic NER. They noted the importance of NER in the context of most natural language processing tasks as well as the paucity of NER software and research in the Arabic language. They suggested integrating two ML techniques: bootstrapping semi-supervised pattern recognition and a supervised CRF classifier. At the time this was published, this combination had been neglected not only in Arabic NER but also in NER in other languages. The exact components of their proposed system were a CRFs classifier, a dual

iterative pattern relation expansion, and the Research and Development International (RDI) Toolkit. The CRFs classifier, a generalisation of the hidden Markov model, was used for segmenting and labelling the sequential data. The dual iterative pattern relation expansion was used as the first pattern extraction algorithm, bootstrapping web pages to find all occurrences of the relation instances in the corpus. Finally, the RDI Toolkit is a combination of the Arabic RDI-ArabMorpho-POS tagger and the RDI-ArabSemanticDB tool. The part-of-speech tagging depends on the word morphology features with high accuracy rates. The RDI-ArabSemanticDB tool utilises an Arabic lexical semantics language resource and the appropriate interface, allowing the processing and storage of more root words with their lexical features. This helps tackle the challenges specific to the Arabic language.

In 2013, Althobaiti *et al.* (2013) presented a semi-supervised algorithm designed to identify NEs in Arabic text. Their algorithm consisted of three components: pattern induction, instance extraction, and instance ranking or selection. Moreover, ASemiNER uses a pattern induction process that infers a set of surface patterns containing seed instances in the training corpus to retrieve all sentences containing each seed. The trigger words are extracted from randomly selected Arabic *Wikipedia* articles based on co-occurrence with the NE. Then, the patterns are generalised, and ASemiNER conducts instance extraction from the training corpus based on what instances match the patterns assigned. The final patterns are matched against the corpus. To prevent confusion with Arabic common nouns, the average NE length of two or three tokens per proper noun are added to the system, stating that increasing the average length of proper nouns to more than two tokens can improve recall but is detrimental to precision and overall result quality. Finally, ASemiNER ranks all the examples according to the number of patterns used to extract them. The NEs with the highest number of distinct patterns are ranked higher than the others. This method of sorting is more appropriate than using frequency of occurrence, as some bad examples are common but are identified by one pattern, whereas some good examples are less common but have more connected patterns. The results are encouraging, matching the result quality of many highly esteemed systems. Eliminating the need for supervision in some areas and identifying a novel method of recognising specific categories of NE, the research by Althobaiti *et al.* can contribute to Arabic NER development.

In 2014, Darwish and Gao (2014) applied SSL via a two-pass method to the extraction of news and microblog data from an ANERcorp news and tweet training set. The use of microblogs presented challenges related to informal language usage, such as abbreviations and short expressions. They applied large gazetteers and a two-pass semi-supervised method as well as domain-appropriate adaptations to navigate the difficulties presented by the tendency to omit data in informal language.

Finally, in 2016, Althobaiti (2016) explored the application of a semi-supervised method combined with a distant learning method to the extraction of person, location, and organisation NEs from

ANERcorp. The elimination of the need for annotated training data and gazetteers meant that the algorithm did not need new data for every change of domain. Althobaiti also suggested that this new approach would improve the usually bad recall of semi-supervised methods up to that point, which is supported by an 8% improvement over the next-best semi-supervised classifier.

2.4.2.3 Unsupervised learning

Clustering is the main approach that is used in unsupervised learning. Based on a similar context, NEs can be gathered from clustered groups. There are different types of unsupervised methods that use lexical resources (e.g., WordNet), lexical patterns, and statistics, which are computed on a large corpus without annotations (Nadeau and Sekine, 2007). Our survey of the literature has not revealed any Arabic NER system employing unsupervised ML yet it has been applied by many NER researchers for other languages. However, the main hurdle is that unsupervised ML requires absolute reliability. Alfonseca and Manandhar (2002) assigned topic signatures to synsets by listing frequent co-occurrences from a large corpus sample. In 2004, Shinyama and Sekine (2004) observed that NEs often appear in sync in various news articles, while common nouns do not. They assessed the punctuality of NEs and their simultaneous appearance to identify rare NEs without supervision. Etzioni *et al.* (2005) used pointwise mutual information and information retrieval to assess the classification of NEs. They found that, using co-occurrence, they could create features for candidate entities and derive many discriminator phrases in an automatic, unsupervised manner.

2.4.3 Hybrid approach

The hybrid method is a combination of the rule-based method and the ML method. Over the past few years, some hybrid systems have been established in order to improve the performance of rule-based and ML systems. Abdallah *et al.* (2012) extended the NERA system developed by Shaalan and Raza (2008) by combining decision trees with the rule-based method. The system recognises three entities, which are person, location, and organisation and has achieved an overall average of 88.87% F-measure. Another system developed by Oudah and Shaalan (2012) combined SVMs and logistic regression and increased the number of NEs from three to 11 types. These are person names, location, organisation, time, measurement, phone number, filename, date, price, percent, and international standard book number (ISBN). They applied this approach to ACE 2003, ACE 2004, and ANERcorp as well as their own corpora. Their system achieved an average of 90% F-measure.

Abdallah *et al.* (2012) utilised a hybrid method to analyse corpora built between ANERcorp and ACE 2003 data extracted from news sources. They applied a J48 decision tree classifier to a rule-based method in extracting person, location, and organisation NEs. The initial experiments showed an

improvement in F-measure between 8% and 14% in comparison with (pure) rule based system and the (pure) machine learning approach.

In 2014, Meselhi *et al.*(2014) presented a new hybrid approach to NER in Arabic. This system was presented with the task of extracting person, location, and organisation NEs from an ANERcorp corpus extracted from newswires and other web sources. The integration of a rule-based approach with a ML approach was combined with the selection and correction of tags to identify any false negatives. Extraction of person entities achieved 96.65% F-measure while the other entities, location and organisation reached 94.8%, and 92.9% F-measure respectively.

2.5 Feature Space for NER

In most NER systems, the input data is transferred word by word into a set of features. These features are used as inputs for the classification phase. Shaalan (2014, p. 482) defined features in the context of NER as “properties or characteristic attributes of words designed for consumption by a computational system”. Features are characteristic qualities of words that have been processed for algorithmic consumption. All words possess vast numbers of individual features, and most research will address multiple features at a time. For example, capitalisation can be assessed as true or false by many simple systems. Features can be abstracted over the text they have been recognised in via feature vector representation, where each word is represented by one or more values. They can be represented by Boolean, numerical, and nominal values (Shaalan, 2014).

Features can be classified into word-level features, list-lookup features, contextual features, and language-specific features (Shaalan, 2014). Word-level features are relevant to the word orthography and structure. Examples of these features are the length of the word, the presence of special markers or characters in the word body, like abbreviation points and hyphens, and the presence of capitalisation in the word gloss in English. List-lookup features are related to the membership of the word in different lists for the purpose of classification of the targeted word. Examples of these features are a stop word list, a gazetteer list, which contains the most frequent entities, and lexical triggers. Contextual features that are related to the targeted word context are defined in this feature set. This includes the type and the words that are adjacent to the targeted word. Table 2.2 summarises the most common features in the literature. A sliding window is usually used in order to determine the boundaries of the analysed context. For instance, if the sliding window size is five, the features of the left two words, the targeted word, and the right two words are considered. The window sizes can differ in different approaches, like ± 1 to ± 3 according to Benajiba *et al.* (2010) and ± 1 according to Benajiba *et al.* (2008). Language-specific features are related to the morphology of the language. They include base phrase chunks and the part of speech along with other morphological features.

Table 2.2 The most common features (Shaalán, 2014)

Category	Features	Description
Word-level features	Special markers	A binary feature indicating the presence of punctuation marks and special characters in a word.
	Word length	A binary feature indicating whether the length of the word is greater than a predefined threshold.
	Capitalisation	A binary feature indicating the existence of capitalization information on the gloss corresponding to the Arabic word.
	Lexical	The surface features of a character n-gram up to a range of characters from 1 to n that indicate prefix and suffix attachment.
List lookup features	Gazetteer	A binary feature indicating the existence of the word in an individual gazetteer.
	Lexical Trigger	A binary feature indicating the existence of the word in the individual lexical trigger list.
	Blacklist	A binary feature indicating the non-existence of the word in an individual blacklist.
	Nationality	A binary feature indicating the existence of the word in the nationality list
Contextual features.	Word n-gram	The features of a sliding window comprising a word n-gram that includes the candidate word, along with preceding and succeeding words.
	Rule-based	The features of a sliding window derived from rule-based NER decisions

Feature selection is a crucial task in NER systems based on ML, as the learning phase depends entirely on the features. The core part of the learning involves the selection of the optimal set features to enhance the performance of the classifier (Benajiba *et al.*, 2008). In the literature, measuring the performance of each feature manually, is the most prominent method used in this regard (Shaalán, 2014). Another method is to make an initial decision of the feature set through isolation testing and then combine this feature set with other features until all features are tested in order to find the optimal feature set. Benajiba *et al.* (2008) used an incremental method, where the effect of each feature is

measured individually, and then the features are ranked decreasingly according to their performance and combined in order to infer the optimal feature set.

In 2009, Benajiba *et al.* later investigated the effect of different feature sets in the three ML systems based in ANERsys. They found that a combination of 15 features resulted in the highest accuracy in terms of performance, suggesting again that the addition of more features disproportionately benefits Arabic NER. Abdul-Hamid and Darwish (2010) tested a simplified feature set to assist in accurately identifying NEs in Arabic. Their goals were to identify simplified features to make the task of Arabic NER easier. Their feature sets were effective despite the simplicity.

2.6 NER Tools and Resources for Arabic Language

This section reviews the available NER tools and resources that are designed for the Arabic language. These include corpora and morphological analysers.

2.6.1 Corpora

Corpora are collections of documents that have been tagged and sometimes pre-processed for language processing tasks. When available, corpora make excellent sources to develop and test Arabic natural language processing systems, and NER systems in particular. However, Arabic language corpora are not as abundant as Arabic digital text nor as abundant to undertake significant research. The scarcity of available Arabic annotated corpora designed for NER has been noted across all domains and has presented a challenge to many researchers who have been forced to build their own corpora to fulfil the requirements of their research. For example, in 2008, Shaalan and Raza (2008) were forced to compose three separate corpora for a cross-domain study. One was a 4 MB reference corpus for person, location, date, time, price, and measurement NEs; another was a 100 KB corpus for company NEs, the last was a corpus for phone number, ISBN, and filename NEs, which had to be composed across various websites to account for the lack of available information. They have acknowledged the lack of diverse corpora and diverse research in Arabic NER and presented their results across diverse corpora, finding their system to be fairly accurate, but the resources ultimately insufficient. Algahtani (2011) also needed to heavily pre-process the corpora used for their NER study. Benajiba and Paulo (2007) had to design and annotate their own corpus for training ANERsys. Similarly, Sawahla and Atwell (2009) had to build their own corpora from 15 Arabic dictionaries to provide control data. It is clear from the reviewed

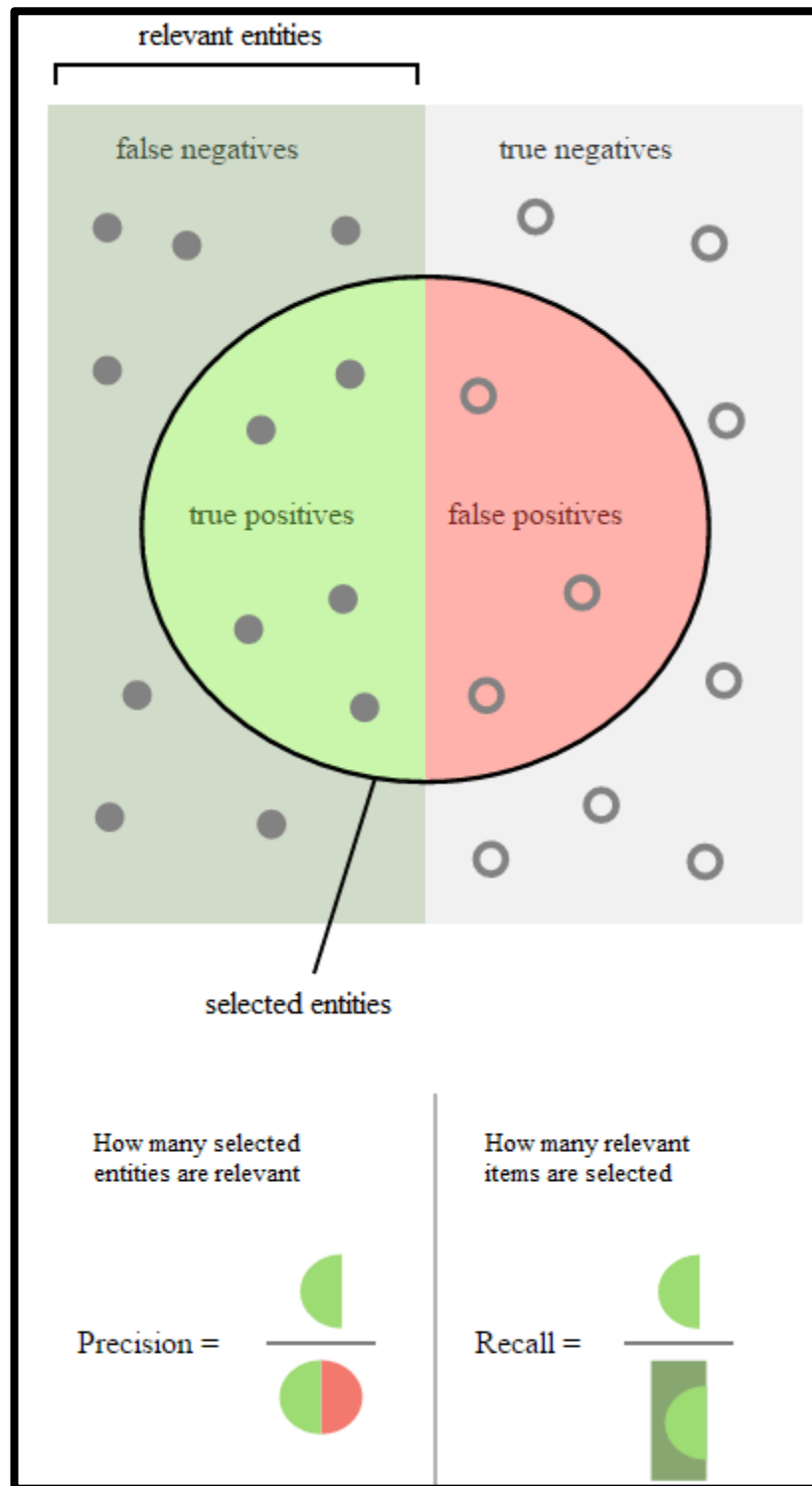


Figure 2.2 A visualisation aid of precision and recall¹.

¹ 'Precision and recall' by Walber with minor alterations by the author, available at <https://commons.wikimedia.org/wiki/File:Precisionrecall.svg> under a Creative Commons Attribution 2.0. Full terms at <http://creativecommons.org/licenses/by/2.0>.

literature that what little annotated corpora are available for Arabic NER tasks are insufficient, and that what is freely available falls vastly short of the demands from the field. This presents a challenging obstacle for Arabic NER researchers.

That said, some corpora are currently available for Arabic natural language processing tasks, and prove invaluable tools in the field. In the literature, there were also some common and more recent examples of corpora in Arabic, which have been and are still being used. Here, some of the corpora that are available are reviewed as follows.

2.6.1.1 ACE

Although only released in 2003, the ACE data sets are one of the earliest data sets developed for use in Arabic natural language processing. The Linguistic and Data Consortium developed and annotated these corpora for automatic content extraction in various languages and was one of the first to cover Arabic. The corpora were updated annually between 2003 and 2005. They are composed of broadcast news data and newswire data. They originally contained 55,000 tokens but increased to 113,000 tokens by 2005 with the addition of Arabic Treebank in 2004 and WebLogs in 2005. The corpora follow a tagset specific to the ACE program and can capture seven coarse-grained and 45 fine-grained NEs. They have been widely used by those with access to them for the purposes of Arabic NER. Despite not having been updated since 2005, they are still regularly used by modern researchers namely by (Benajiba *et al.*, 2009; Abdallah *et al.*, 2012). These corpora are, however, inaccessible to the public.

2.6.1.2 ANERcorp

The ANERcorp is the first freely available annotated corpus specifically for Arabic NE extraction. Developed by Benajiba *et al.* (2007), it was based on the structures of the publicly unavailable ACE corpora as a more accessible corpus. Four categories of NE can be extracted: person, location, organisation, and miscellaneous. It is newswire based and manually annotated for accuracy, containing 150,000 tokens. Moreover, 11% of the tokens are NEs, of which 39% are person NEs, 30.4% are location NEs, 20.6% are organisation NEs, and the remaining 10% are miscellaneous. It is now considered a standard data set for the evaluation and comparison of Arabic NER systems and is one of the benchmarks used for system performance. Its free accessibility and comprehensive data are the main reasons for its popularity as an evaluation benchmark. However, as more research delves into domain-specific NER, the value of the four categories used by ANERcorp may lessen compared to corpora specific to the domain.

2.6.1.3 AQMAR

The AQMAR corpus is one of the most recent Arabic NER corpora that is also freely available. It was developed by Mohit *et al.* (2012) as part of the American and Qatari Modelling of Arabic Project. It consists of Arabic *Wikipedia* articles, which have been manually annotated for NEs in four categories (person, organisation, location, and miscellaneous). Branching out from the early corpora based on news domains has allowed this corpus to be based on other, more diverse domains. Further, AQMAR2, based on 28 Arabic *Wikipedia* articles, contains 74,000 tokens from the domains of history, science, technology, and sports. The AQMAR corpus is a valuable contribution to the rapidly developing diversity of Arabic NER.

2.6.1.4 WikiFANE_{Selective} and WikiFANE_{Whole}

In a similar manner, Alotaibi and Lee (2013) developed the WikiFANE corpora based on Arabic *Wikipedia* articles. These corpora are manually annotated but were collected using the ACE NE taxonomy. However, there is a major modification in that the NE class of person is subdivided into nine fine-grained classes: artist, athlete, businessperson, engineer, police, politician, religious, scientist, etc. The coarse-grained class of product NE is also added. WikiFANE is divided into two corpora: WikiFANE_{Whole} and WikiFANE_{Selective}. WikiFANE_{Whole} contains all the sentences retrieved from the scanned articles and comprises 2,023,496 tokens, whereas the Selective corpus contains only those sentences containing at least one NE phrase, reducing the token count by over two thousand to 2,021,177.

2.6.1.5 NewsFANE_{Gold} and WikiFANE_{Gold}

Alotaibi and Lee (2014) went on to develop NewsFANE Gold and WikiFANE Gold in 2014 using a similar process and the two-level taxonomy employed in the development of the previous two WikiFANE corpora. NewsFANE is a newswire corpus making use of the same textual data in ANERcorp, only annotated for more fine-grained NER. WikiFANE Gold, like the other WikiFANE corpora, draws from Arabic *Wikipedia* articles; however, it has only a quarter of the tokens at 500,000. That said, both are more fine-grained than previous corpora.

2.6.2 Morphological analysers

As Arabic presents a complex morphology, it can often be a challenging language for natural language processing systems. However, with the increase in available Arabic digital texts and corpora has come a strong interest in the applications of natural language processing systems for Arabic language. This has required the development of additional tools designed to handle the morphological differences

between Arabic and other popular natural language processing languages, such as English or Chinese. The following is a list of different morphological analysers and other text pre-processing tools that are the most widely used in Arabic NER.

2.6.2.1 BAMA

A recurring feature in the literature, BAMA is one of the most widely used tools when it comes to Arabic natural language processing, such as the work by Buckwalter (2002), Farber *et al.* (2008), Elsebai *et al.*, (2009), Elsebai and Meziane (2011), and Al-Jumaily *et al.* (2012). Moreover, BAMA files its Arabic-English lexicon into prefixes, suffixes, and stems, with 299,618 and 82,158 entries, respectively. These are then supplemented with three morphological compatibility tables that control the prefix-stem combinations, the stem-suffix combinations, and the prefix-suffix combinations, representing 1,648, 1,285, and 598 entries, respectively. Additionally, BAMA can be accessed through the Linguistic Data Consortium.

2.6.2.2 MADA+TOKAN

The Morphological Analysis and Disambiguation for Arabic (MADA) tool is invaluable to Arabic NER. Built on top of BAMA by Habash *et al.* (2009), MADA+TOKAN is a two-part tool that disambiguates Arabic NEs before tokenising them. The morphological analysis that MADA conducts can then be moved into the process of tokenising the isolated stem deterministically. The TOKAN component of the system allows the resulting disambiguated analysis to be tokenised via any specified tokenisation scheme. This is important because there are many different ways to tokenise the Arabic language, and deciding which one is most appropriate can depend on the context of the research.

This tool is essential to the broadening spectrum of Arabic natural language processing research, providing one basic solution to all the core problems identified in Arabic natural language processing. It handles tokenisation, diacritisation, morphological disambiguation, part-of-speech tagging, stemming, and lemmatisation. This helps to break down and manage Arabic texts that would otherwise be too complex for NER software to identify NEs with much accuracy. This tool was widely used in the literature by many researchers such as (Farber *et al.* 2008; Benajiba and Rosso 2008; Oudah and Shaalan 2012; Oudah and Shaalan 2013).

2.6.2.3 AMIRA

AMIRA is a statistical toolkit designed to process Arabic morphology. In 2009, Diab presented AMIRA as a solution to the problem of processing modern standard Arabic. AMIRA addresses many different aspects of natural language processing in Arabic and has many functions, all of which can

contribute to NER in Arabic. The tokenisation process of AMIRA allows the user to select token count as well as the option of clitic tokenisation, where conjunctions, prepositions, pronouns, future marker clitics, and definite articles are separated as well as prefixes and suffixes. This provides the added benefit that the root of each word can be located for improved NER, a highly important task in Arabic. To do this, AMIRA applies a chunking scheme on the character level and locates a chunk boundary. Using an IOB (inside/outside/beginning) annotation scheme, the characters are selected as being inside, at the beginning, or outside a chunk. For inside and beginning characters, there are five classes. The goal is to produce a text where the word tokens are all words as represented in a standard Arabic dictionary. In addition, AMIRA utilises part-of-speech tagging, including an optional standard tag set of 25 tags and an extended tag set of 72. The part-of-speech tagging by AMIRA adopts an approach to classification based on SVM, and both taggers have an accuracy over 96%. This improves the results of later, higher processes, such as base phrase chunking. In AMIRA, the model is designed to produce the longest possible base phrases, while avoiding internal recursion. The breadth of applications of AMIRA as well as its high adaptation to the Arabic language and high accuracy and flexibility make it a useful tool for NER in new and underexplored domains.

2.6.2.4 MADAMIRA

With the best components of both MADA and AMIRA, MADAMIRA is the end result of combining the two strongest commonly used systems in Arabic natural language processing. Moreover, MADAMIRA makes use of a streamlined Java implementation, which results in a more resilient, malleable, and fast process. Developed by Pasha *et al.* (2014), it efficiently handles modern standard Arabic as well as Egyptian Arabic, illustrating how few alterations are necessary for MADAMIRA to handle the dialectal variations of the Arabic language. In addition to improving performance and duplicating the functions of MADA and AMIRA tools, MADAMIRA “was designed to be fast, extensible, easy to use and maintain” (Pasha *et al.*, 2014, p. 1095). Moreover, MADAMIRA maintenance is easier and unlike AMIRA, users do not need to install any other third-party software in order to run the tool. Another additional feature which was not present in MADA or AMIRA, is the support of XML and HTTP where the input and output text could be provided as plain text or in XML format (Pasha *et al.*, 2014).

2.7 Conclusion

In 2007, Nadeau and Satoshi (2007) carried out a survey of the 15 years of research in the NER and classification fields, ranging from 1991 to 2006, revealing that earlier systems used hand-written rule-based algorithms far more often than the more recent systems which were increasingly leaning towards

ML techniques. This shift is a result of the explosion of research in developing new and better ways of employing ML in NER.

The domains of NER and classification research had not been expanded greatly, with most studies using the same domains for their investigations. This means that many domains, such as medicine, cannot be wholly addressed by a pre-existing system, as domain dependence interferes. Combined with language barriers, this is a problem that is only slowly being addressed by the adaptation of various systems and toolkits to specific document types. This also means that certain types of NE such disease and symptoms are not adequately addressed, due to the lack of research in the domains where they have a strong presence. Earlier work focused on extracting proper nouns in general (e.g. named of people, location, and organization) but lately much research is focusing on more complex NE types.

The majority of NER and classification work that was done focused on English, but with a growing representation from German, Spanish, Dutch, Japanese, Chinese, French, Greek, and Italian. However, Arabic NER and classification was only just starting to develop large NER projects.

However, there are not many analysers or corpora available for Arabic NER study as this is a developing research area. In time it is likely that the available tools and corpora will be as abundant and successful as those used in English, Chinese, and German.

In conclusion, NER in Arabic has made much progress; however, the field remains underexplored, especially considering the global effects of the Arabic language. In many ways, becoming a popular NER language around the same time as many other languages were moving from hand-written systems to ML systems has affected Arabic NER research. Especially when we consider the complex morphology of Arabic, it becomes clear that much work still needs to be done to help bring the field to the level of other significant languages in terms of NER. When it comes to the variety of domains and accessible annotated corpora, Arabic NER trails even further behind, possessing large data gaps that must be filled to encourage further progress. Therefore, this study intended to contribute to the rapidly growing body of Arabic NER research and assist in bringing Arabic NER to the foreground.

In this chapter, we surveyed the current research efforts in the Arabic named entity recognition field including the NER approaches, the evaluation metrics, and the NER tools and resources for Arabic language. In the next chapter, we will discuss the theoretical foundations that underpin our research which are the natural language levels and Bayesian belief networks.

Chapter 3: Theoretical Foundations

3.1 Introduction

This chapter reviews the theoretical foundations that underpin this research. As this research project analyses texts to recognise specific named entities, the relevant literature related to natural language processing (NLP) is covered in the first section. Bayesian belief networks (BBNs), which are applied to represent and learn from the knowledge acquired from the NLP levels in order to recognise the entities, are introduced in the second section (Figure 3.1).

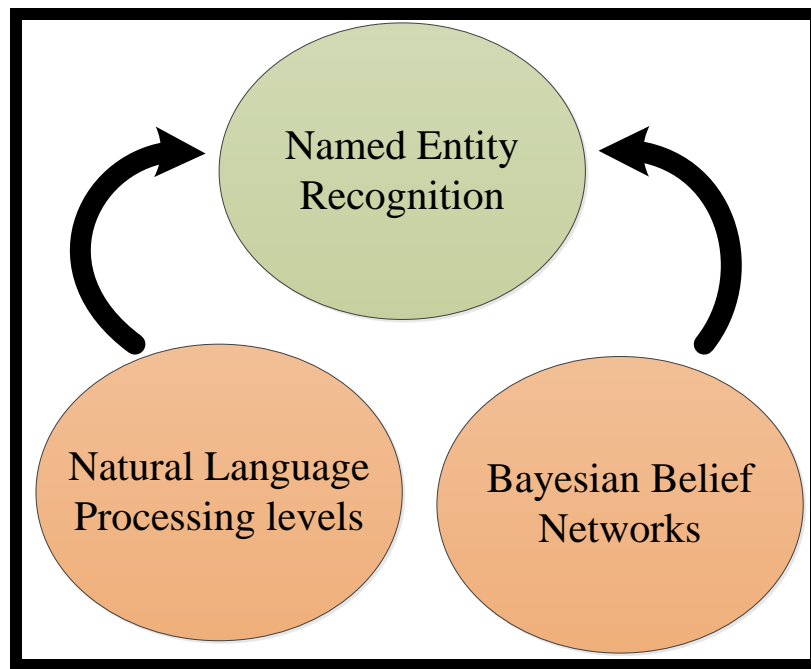


Figure 3.1 Theories that underpin this research.

3.2 Natural Language Processing Levels

Human analysis is generally, if not always, the most favourable form of text analysis. After all, humans can more accurately understand language, making exceptions known and even incorporating new terms, than a machine can. However, the human hours it would take to process large corpora manually would be extraordinary and not a good return on investment. Steedman noted that “computer science provides a rich source of models for theories of all three modules of the human processor”, as “within the Artificial Intelligence paradigm [one can find] both theories and working examples of the way in which syntactic processing, semantic processing, and referential processing can be interleaved and may in very restricted senses interact during processing” (Steedman, 1994, p. 231). However, although he

viewed this in some ways as an advantage, Steedman was quick to acknowledge that “programming languages are strikingly unlike their natural counterparts” (Steedman, 1994, p. 231). This is very well when it comes to the areas where computational analyses are more efficient, but what about the many ways in which language is a variable object? This is where NLP can be beneficial.

Natural language processing (NLP) is aimed at facilitating computers to understand and manipulate natural language text or speech by exploring how humans understand and use natural languages (Chowdhury, 2003). People extract meaning from natural language text or speech using seven different levels: phonetic, morphological, lexical, syntactic, semantic, discourse, and pragmatic (Feldman, 1999; Liddy, 1998).

In computational linguistics, when a text or even a word is processed by a program, this program must be calibrated to understand the nuances of meaning behind the text it has been given. This process of mining, when not calibrated to natural language data, could result in lost or misrepresented information. To a human, it is natural to break down various forms of meaning and understand them, sometimes in a matter of seconds. This is very helpful when analysing small text samples, but when analysing large corpora, this would be a very time consuming process. Through NLP, we can attempt to recreate that natural human understanding of language in our programs, rendering more accurate data from text mining (Popowich, 2005). The six levels of NLP are described below. However, the phonetic level is not discussed because this research focus on only texts not speech.

- Morphological level

Morphology involves the study of the shapes words take, as text and as sounds, and the ways in which words across a language bear relation to each other. As words are a combination of form and meaning, it is important that any program designed to analyse words ‘understands’ not only the form these words take but also how they are related to the text and to the source language in general. For example, the addition of prefixes and suffixes may create words that differ in form but are almost identical or related in meaning, (the English addition of -s to create a plural or the Arabic addition of the *Al* determiner to create a definite form). Likewise, many nouns can be transformed into adjectives or verbs, to make them descriptive words or action words that, while bearing some relation to the original noun, have an entirely different meaning (Anderson, 2003). For a program to correctly analyse corpora, therefore, it must come equipped with some ability to categorise words based on meaning and context. At the morphological level, NLP involves the study of the component structure of a word (Liddy, 1998). In order to understand an unknown word, humans break it down into its component parts and try to understand the parts. Likewise, any NLP system can do the same. For example, in Arabic, words consist of prefixes, a stem or a root, as well as suffixes with different combinations. If an NLP system

cannot find the meaning of the whole word, it can extract its root by breaking it down (Liddy, 2001). The morphological analysis level is relevant to our work, as one of the pre-processing steps of our system is text tokenisation. At the tokenisation step, a morphological analysis is carried out where words are broken down into prefixes, stems, and suffixes. For instance, a sentence in English such as "and they will write it" can be split into five tokens, while in Arabic this is expressed in one word وسيكتبونها (wsyktbonha). As this example shows, the conjunction "and" and the future marker "will" are represented as prefixes by the letter و and س respectively, while the pronouns "they" and "it" are represented by the suffixes ون and ها respectively.

- Lexical level

The lexicon of a language is the collection of words that the language comprises. Lexical proficiency is an important part of understanding language. To be lexically proficient, a human or program must not only know many language features but also understand each feature well and be able to process each feature quickly and on demand. For a program, there are many hurdles to understanding a lexicon that a human may not consider. These are the qualities that any given language feature possesses. Polysemy is the number of meanings related by extension a word has. For example, the English word 'head' could refer to the body part or to a president. In the same way, the Arabic word "عين" could refer to the body part or to a spy. Hypernymy is the specificity of a word. For example, pigeon, duck, and seagull are all hyponyms of bird (their hypernym), while semantic co-referentiality assesses words relative to semantic similarity. Word frequency measures how often the words occur, and word concreteness differentiates between specific and abstract words. Word familiarity represents the cultural value of a word, and word imaginability explains how easily a word evokes an image. Word meaningfulness demonstrates how closely associated words are to other words, and word length shows the total number of characters in a language feature. All of these could be the target of research, and many of them need to be considered even if they are not the main objective of the research. Different kinds of processing can contribute to lexical understanding. One of these methods is to label each individual word with part of speech (POS) tags. Using a lexicon could also be required for lexical level analysis (Liddy, 2001). The lexical analysis level is important to our research as POS tagging is performed at the pre-processing step of our system. Assigning a POS tag to each token in our corpus can help our BBN understand the meaning of these individual tokens. Also, at feature extraction step, different lexical lists (lexical markers list- Gazetteers, Stopwords list) were produced during this research to help the BBN 'understand' the general meaning of these tokens.

- Syntactic level

Syntax covers the assessment of correct grammar and how we reach the conclusion and consensus that grammar rules are applied appropriately. However standard formal grammar may not adequately explain the meanings conveyed in a text or corpora, especially not when the document is legal or technical and may have a different way of ordering its components. The context and detailed properties of the words are key to an accurate reading of the syntax. Any program analysing syntax needs to be adequately prepared to understand these variables in the context of the text or corpora being analysed (Schubert, 2015).

The focus is on the grammatical structure of the sentence and on the position and type of each word, this usually requires parsers (Liddy, 2001). In our work, a shallow syntactic analysis is performed in order to extract appropriate patterns that characterise our corpus. Details of these patterns are presented in Section 4.4.3. The POS tags of a $-/+2$ words sliding window (5 words) are also applied in our BBN in order to detect any syntactic pattern. A deeper syntactic analysis can be done using a base phrase-chunking tool. However, too much syntactic depth is not required, as the interpretation of a complex extract from a text is determined by the interpretation of its components and syntax largely assists in identifying the semantic interpretation of the same extracted elements (Steedman, 1994).

- Semantic level

Semantics is the study of the meaning of words, phrases, and symbol expressions in NLP. Computational semantics combines an understanding of formal semantics with computational linguistics and automated reasoning. The collection of semantics can be performed alongside a syntactic analysis with little to no loss of information (Steedman, 1994). Semantic analysis varies in process, depending on the end goal of the research. First-order logic is generally seen as the favoured starting point. This is because first-order theorem provers are currently developed enough to offer insight into semantic reasoning and because first-order logic can process a wide variety of phenomena. The system will break text down into language formulas, replacing logical variables with representations, leaving only linguistic constants at the end. The model then interprets the resulting formula using variable assignment functions. However, it is important to not use this formula rigidly. Semanticists need to use ‘natural language metaphysics’ (Bach, 1986) to account for the fact that human language can be highly variable (Blackburn and Bos, 2003). According to Gabrilovich and Markovitch (2007, p. 1), representation of natural language semantics requires “access to vast amounts of common sense and domain-specific world knowledge”. There are lexical databases available, including WordNet, and thesauri that allow for encoding of the relations between lexical items in order to solve the semantic disambiguation of words. However, even with these tools, word sense

disambiguation remains impossible in many instances (Gabrilovich and Markovitch, 2007). In our study, every token in our corpus is labelled with an appropriate tag representing a shallow semantic meaning of those tokens that determines whether these tokens are specific named entities or not.

- Discourse level

Discourse information is very important in language production and analysis, among other linguistic processes. Defining discourse as a separate entity from pragmatics has often been a point of contention. Discourse used to be used primarily considered part of pragmatics, focusing on discourse factors and the behaviour of the speaker and recipient, namely, how word and expression choices are influenced by culture and, in turn, create a culture. However, (Dijk, 1983) has illustrated that discourse is a valuable and independent part of the system of language. Using a Bayesian behavioural model, the addition of discourse knowledge assists in recovering categories and sorting tokens more accurately. More specifically, pronouns were found to be categorised with greater accuracy once discourse was considered (Orita, 2015). At the syntactic and semantic levels, NLP deals with sentence-length units. Discourse analysis deal with units that are longer than sentences. It “focuses on the properties of the text as a whole that convey meaning by making connections between component sentences” (Liddy, 2001, p. 9). Anaphora resolution and discourse structure recognition are examples of discourse analyses.

- Pragmatic level

Pragmatics is primarily concerned with inference, which is the information a reader or listener gains without being explicitly told. It is, in short, deductive reasoning derived from formal elements of language structure, such as semantics, syntax, and lexicon, combined with cultural elements of language, such as discourse and context. As an organic human trait, this can be a difficult field for computers to navigate. It covers everything from the previously discussed discourse to the interpretation of speech acts and an understanding of coherence. Pragmatics can be covered by logic-based approaches and probabilistic approaches, but, due to the depth of meaning and variability of pragmatics, a combination of both generally delivers the best results (Jurafsky, 2003). The pragmatic level includes “understanding the purposeful use of language in situations, particularly those aspects of language which require world knowledge” (Liddy, 1998, p. 14). This knowledge comes from the outside world, so it is from outside the content of the document. It could include intentions, goals, and plans.

3.3 Bayesian Belief Network

The BBN is a probabilistic graphical model which is also known as a recursive graphical model, belief network, casual probabilistic network, casual network, influence diagram, and several other designations. When discussing BBNs, many authors will also have a slightly different interpretation of the model, which can further confuse understanding. The main agreement is that BBN breaks down probabilities of outcomes based on a series of reliable equations, providing a graphical structure with a set of probabilities (Daly *et al.*, 2011). For the purposes of our research, 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 (Wooldridge, 2003). Such Belief networks are able to represent probabilities over any discrete sample space so that the probability of any specific sample point from that space can be computed from the probabilities outlined in the BBN (Daly *et al.*, 2011).

The BBN is an approach that allows the user to form a hypothesis H about the world based on the given evidence e :

$$p(H|e) = p(e|H)p(H) / p(e) \quad \text{Equation 3.1}$$

where $p(H|e)$ is sometimes called the posterior probability, $p(H)$ is called the prior probability, $p(e|H)$ is called the likelihood of the evidence (data), and $p(e)$ is a normalising constant (Heckerman, 1998).

A famous example of a Bayesian network is presented by Professor Judea Pearl from the University of California, Los Angeles. The example is as follows:

“ ‘I’m at work, neighbour John calls to say my alarm is ringing, but neighbour Mary doesn’t call. Sometimes it’s set off by minor earthquakes. Is there a burglar?’ ”

The variables of the network are: burglar (B), earthquake (E), alarm (A), JohnCalls (J), and MaryCalls (M). The network structure is presented in Figure 3.2. It consists of five nodes where each node represents a variable. The network topology reflects the following ‘causal’ knowledge:

- A burglar can set the alarm off.
- An earthquake can set the alarm off.
- The alarm can cause Mary to call.

- The alarm can cause John to call.

Each node in the network has a conditional probability table (CPT), which indicates the probability of the node being true given its parent nodes' value.

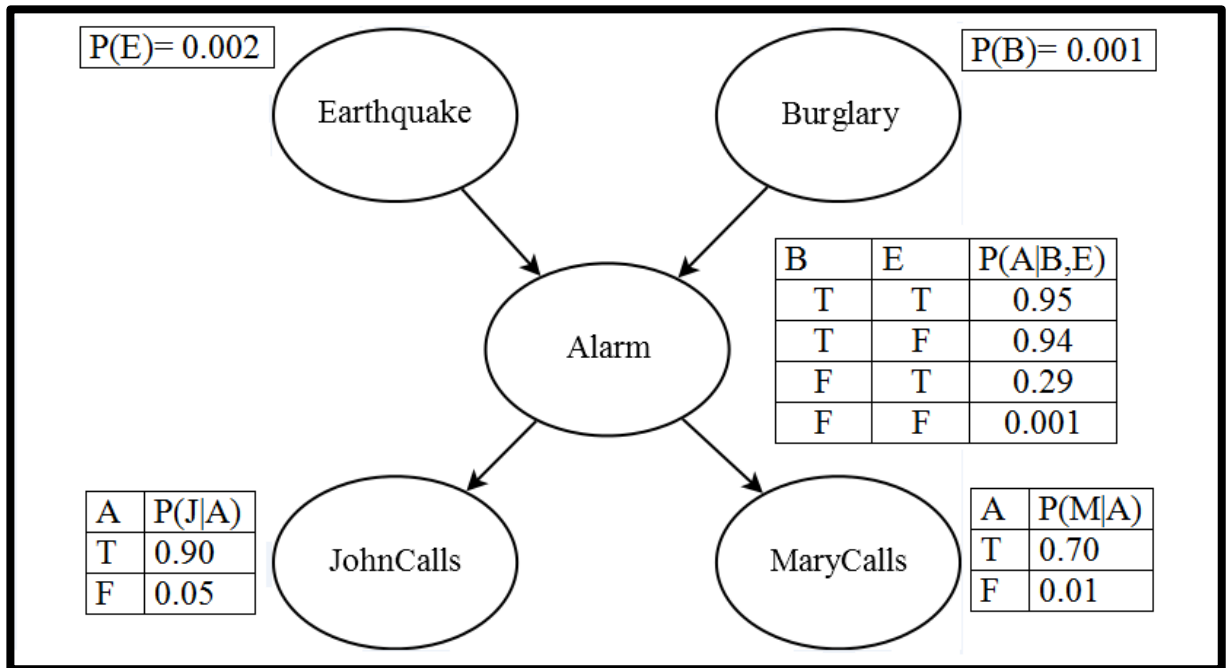


Figure 3.2 An example of a typical Bayesian network.

Given the above example, the Bayesian network gives us the chance to model any scenario and compute its probability. Examples of these scenarios are:

- The probability of burglary given that John and Mary call.
- The probability of Mary calling given that John did not call and there is an earthquake.
- The probability of the alarm going off given John is calling.

The use of belief networks has become widespread partly because of their intuitive appeal. Practical uses for Bayesian networks were outlined in 1998 by Niedermayer, and many of them still hold true to this day. If anything, they have been further explored and expanded. The speed with which BBNs have been adapted to these various fields shows both their variety of potential as well as how easily they

transform and apply to any given field. At that time, the National Aeronautics and Space Administration (NASA) was heavily investing in Bayesian research. This was because deep space exploration could result in novel scenarios playing out, and they needed to work out the most likely results of different actions. Adding more and more data, they were able to predict events that even their top researchers would not have devised. This shows that Bayesian networks have distinct advantages over human intuition, while at the same time working with our intuition and knowledge (Niedermayer, 1998). These same developments have been worked with and expanded in numerous fields of research since then.

BBNs have been deployed successfully in many applications. They have been used for military applications and, more specifically, in threat evaluation by Johansson and Falkman (2008) who developed a threat evaluation system that could handle data better than imperfect observations of humans and other systems. Zou and Conzen (2005) adapted a dynamic Bayesian network (DBN) for the prediction of the gene regulatory system from time course expression data. They presented the DBN-based approach and revealed that it had increased accuracy and reduced computational time compared to other DBN approaches, improving the process of prediction. Their new approach limited potential regulators to genes with early and simultaneous expression changes, effectively allowing the BBN to learn as it processed the information. This resulted in a limited number of potential regulators and a smaller search space. They also employed lag estimation to further increase the accuracy of predicting gene regulatory networks. Their results demonstrate that the approach can predict regulatory networks with greater accuracy and less computational time.

Ticehurst *et al.* (2007) have also used BBNs in the assessment of the sustainability of eight coastal lake-catchment systems, which are located on the coast of New South Wales in Australia. They found that BBNs factored all the variables in a less complicated and easier to understand way than other processes. Presenting a case study as evidence, they proposed that BBNs should be used more often for ecological analysis, due to their simplicity of use and reliable output.

The BBN has been applied in the medical domain. Nikoyski (2000) found that Bayesian networks assisted in numerical probabilistic analysis when the information provided was incomplete or only partially correct. This occurs often when studies publish indirect statistics. Although nothing can replace the actual information, a BBN is as close as it gets to filling in the blanks. The techniques were discussed in the practical contexts of designing diagnosis devices.

Velikova *et al.* (2014) also applied BBN in healthcare. They noted that although it is still difficult to bridge the gaps between Bayesian networks, they are still the best technology available for modelling medical problems, including personalisation of healthcare. Inputting knowledge of diseases based on

the interpretation of patient data allows the BBN to predict the progression of the disease. They used preeclampsia to illustrate the use of this model.

However, to our knowledge, there has been no research using BBNs to extract named entities, yet it has been successfully used for NLP of the English language, such as spell checking (Haug *et al.*, 2001), text categorisation and retrieval (Yang, 1994), and speech recognition (Bilmes, 2004). This research aims to demonstrate that BBNs can provide an efficient and novel approach to extracting named entities from Arabic medical texts.

One of the main advantages of belief networks is that they concisely represent probabilistic relationships. To this end, we should only factor in the known dependencies instead of assuming that all variables are dependent on other variables. This puts the researcher in a good position to acquire and represent knowledge in their domain. For this reason, belief networks have been favoured for data interpretation (Cooper, 1990). Bayesian networks give researchers the material required to put their domain expert knowledge to use in the discovery process. This is significant, as other techniques rely more on coded data, meaning BBNs at worst will provide additional data and at best will provide the exact data required. The BBNs are also easier to read and understand due to the use of nodes and arrows. Using nodes and arrows to signal variables of interest and illustrate the relationships of variables gives researchers room to encode their personal expert knowledge on any given domain using graphical diagrams, which makes it much easier to understand and analyse the output of the BBN. Bayesian networks are designed to make the most of encoded knowledge and to increase the efficiency of their modelling process, which then goes on to improve the accuracy of their predictions. The BBNs are also noted to be at an advantage when detecting and capturing interactions among input variables. Decision trees or CART sometimes may seem to produce a more accurate classification, but this is because they only factor in the relationship between output and input variables. However, even at the expense of the accuracy of classification, the ability of BBNs to also capture the relationships between input variables adds value. The BBNs are still fairly accurate in their predictions, even when using incomplete data. This also means that BBNs are less influenced by small sample sizes (Lee and Abbott, 2003). A few disadvantages do exist. For example, some researchers note there is a distinct absence of commercially available BBN learning algorithms and that the computational methods required are highly complicated (Lee and Abbott, 2003), but these do not detract from belief networks' value when used correctly.

During the conduction of this research, the relevance tree algorithm, which is an exact probabilistic inference algorithm for Bayesian networks and dynamic Bayesian networks, is used to perform the learning and inference. The relevance tree algorithm is based on the well-known "Relevance reasoning" approach. Relevance reasoning is "the process of eliminating nodes in the Bayesian

network that are unnecessary for the computation at hand” (Jammalamadaka, 2004, p.34). Relevance reasoning is based on the d-separation; it can be used to improve the efficiency of Bayesian belief network by identifying and pruning irrelevant parts of a network (Lin and Druzdzel, 1998). The aim is to identify those target nodes involved in the inference task, and to eliminate irrelevant nodes which are not likely to affect the computation on those target nodes (Jammalamadaka, 2004). The elimination process involves three steps:

- The elimination of computationally unrelated nodes. This is the most important step and is based on the d-separation criterion. Nodes that are d-separated from target nodes are probabilistically independent and can be removed; the number of nodes are then reduced.
- The elimination of barren nodes (i.e. nodes with no descendants or no evidence). Although barren nodes may depend on the evidence, they have no impact on the probabilities at the target nodes so they are computationally irrelevant.
- The elimination of nuisance nodes. A nuisance node is computationally related to a target node given a certain evidence (i.e. is not d-separated). The elimination process of the nuisance nodes is carried out by marginalising these nodes into their children (Jammalamadaka, 2004).

3.4 Conclusion

This chapter has presented an overview of the theoretical foundations that underpin this research, which are the NLP and BBN. The NLP levels underpin the first stage of our study, which is described in Chapter 4, where the textual data is pre-processed and analysed and describing how features are extracted. BBNs underpin the second stage, which is described in Chapter 5, describing how the BBN is used to learn from the annotated data and features in order to predict and recognise the entities.

As has been observed, these multiple levels of Natural Language Processing, have their advantages and disadvantages when it comes to machine processing as opposed to human analysis. It is important to note that though computational processing is continually improving, it is still impossible to teach computer systems to understand these levels in the same way as humans. This is mediated by passing some morphological, lexical, syntactic and semantic analysis to our Bayesian Belief Network. Morphological analysers are used to break down words and extract the meaning from its base components. This is essential due to the agglutinative nature of Arabic, as well as the complexity of medical terminology. The lexical level applies Part-of-Speech tagging necessary at the pre-processing

stage of our medical corpus. This allows our BBN to extract the meaning of individual lexical tokens. A shallow syntactic analysis is performed to extract patterns that are representative of our corpus. A shallow semantic tagging of the tokens in our corpus is used to determine whether or not the token is a relevant named entity. The application of available NLP tools allows us to better simulate the ways in which humans extract meanings from text.

Bayesian Belief Networks can help break down the probability of certain outcomes based on the extracted features. This approach is a useful contribution to our research as it allows us to adjust the network structure for better processing, using our own expert knowledge to make the system more accurate. However, reality is far more nuanced than any extant system can account for. Real language is far more complicated than can currently be understood by any system.

Chapter 4: Named Entity Recognition with NAMERAMA

4.1 Introduction

This chapter describes the process of implementing the NER system in the Arabic medical domain, referred to as NAMED Entity Recognition Approach for Modern Arabic (NAMERAMA) in this thesis. The implementation is described in two chapters: this chapter focuses on the natural language processing stage, whereas Chapter 5 describes our BBN approach to the classification and recognition of appropriate named entities. Chapter 4 starts with an overview of the NAMERAMA system, describing its architecture components in Section 4.2. Due to the lack of an Arabic NER corpus related to the medical domain, we have built our own corpus, which consists of 27 Arabic articles related to the cancer domain and is discussed in Section 4.3. The processing of the Arabic corpus is outlined in Section 4.4.

4.2 System Overview of NAMERAMA

NAMERAMA is developed to recognise and extract specific named entities and consists of two main stages. The first stage, which relates to Arabic language processing, comprises three main steps: pre-processing, data analysis, and feature extraction, whereas the second stage focuses on the application of the BBN to classify, recognise, and extract the relevant named entities. Our system architecture is described in Figure 4.1, where each stage is coloured differently. At the pre-processing steps, data tokenisation and POS tagging are carried out using the AMIRA tool; this is described in Section 4.4.1.1. The output is checked and corrected manually by the researcher after data tokenisation and then is annotated. At the data analysis step, frequency, collocation, and concordance analyses are carried out to extract the optimal feature set to be used to train the classifier at the second stage. The analysis of the corpus had yielded a set of features, namely the gazetteers, the appropriate lexical markers, a set of linguistic expressions (patterns), and a list of stop words list. These extracted features are then ready to be converted into a matrix vector to be trained by our BBN stage in order to extract relevant named entities for our medical corpus.

4.3 Corpus Description

The medical corpus used to illustrate our research approach is extracted from the King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE) website. The KAAHE was initiated through the collaboration between the King Saud Bin Abdulaziz University for Health Sciences and the Saudi Association for Health Informatics and was further developed by the National Guard Health Affairs, the Health on the Net Foundation, and the World Health Organisation. The KAAHE content was provided by the U.K. National Health Services (NHS Choices). The KAAHE became the official health encyclopaedia in May 2012 (Saudi E-health Organisation, 2012). The KAAHE is a reliable health information source, containing abundant information written in modern Arabic, which is in easily understandable language and is appropriate for users from various community groups (Alsughayr, 2013). We have extracted 27 articles describing different types of cancer, providing the study with an initial total of 50,256 tokens. After tokenising the articles, the number of tokens has increased to 62,504 due to the linguistic structure of the Arabic language where pronouns and particles such as conjunctions, prepositions, and determiner are agglutinated to the lexical item and must be segmented off for further processing. As a result, the size of our corpus has increased by around 24.37%, as shown in Table 4.1. Figure 4.2 shows a sample of the untokenised text in Arabic, its translation into English, and its tokenised version.

4.4 Arabic Language Processing Component

The first stage of our NAMERAMA system is based on a pipeline process where three main steps are carried out sequentially: the data pre-processing step, data analysis step, and feature extraction step.

4.4.1 Data pre-processing steps

The data pre-processing step is essential to any successful NER system. It aims at preparing the data for further exploitation by the next steps and consists of three main steps, namely data tokenisation, POS tagging, and data annotation described below.

4.4.1.1 Data tokenisation with AMIRA

The data tokenisation is carried out using AMIRA, a set of tools developed at Stanford University. It includes a tokeniser, a part of speech tagger (POS), and a base phrase chunker providing a shallow syntactic parser. It is based on supervised learning and uses Support Vector Machine (SVM) to perform the tokenisation of Arabic words (Diab, 2009; Diab *et al.*, 2007). AMIRA allows the user to determine the tokenisation scheme among schemes such as those described below.

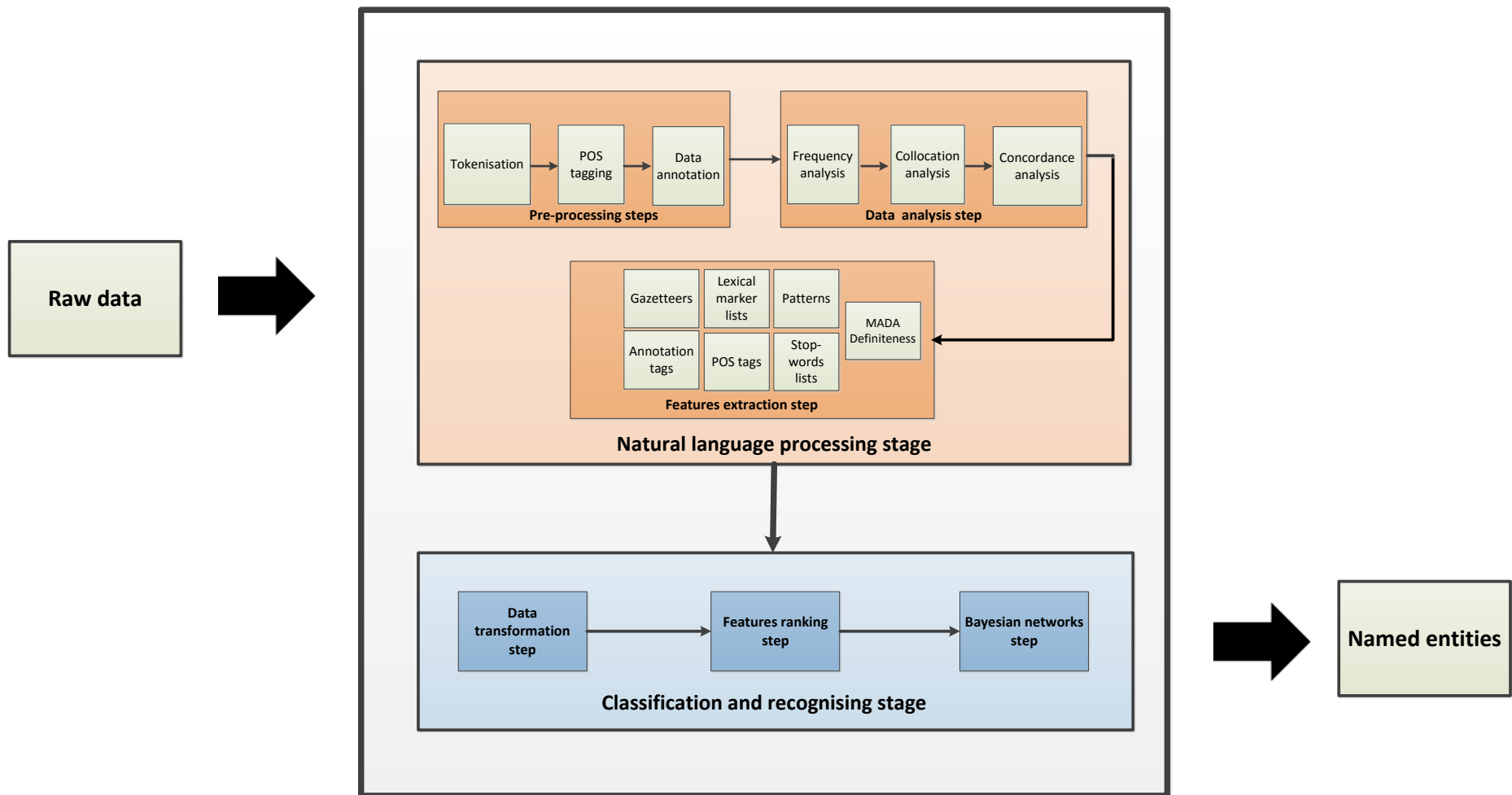


Figure 4.1 The architecture of NAMERA system.

هناك نوعان رئيسيان لأورام الدماغ وهما: الورم الدماغي الأولي والورم المنتقل. يبدأ الورم الأولي في الدماغ، أمّا الورم المنتقل فيبدأ في مكان آخر من الجسم، ثمّ ينتقل إلى الدماغ. هناك نوعان من الأورام الأوليّة وهي: الأورام الحميدة والأورام الخبيثة. ولا تحوي الأورام الحميدة خلايا سرطانية، أمّا الأورام الخبيثة فإنّها تحوي خلايا سرطانية. تُسمّى أورام الدماغ الأوليّة الحميدة الأكثر شيوعاً الأورام السحائيّة. وهي تبدأ في غطاء الدماغ الذي يُسمّى الجافية. وهي أكثر شيوعاً لدى النساء منها لدى الرجال. أمّا في المرضى الكبار في السن، فيجب مراقبة الأورام السحائيّة الصغيرة عند عدم وجود أعراض مهمّة. قد يحتاج الأمر إلى إجراء عملية جراحية لاستئصال الورم السحائي الأكبر حجماً، ومن غير المحتمل أن يعود ظهور الورم السحائي عند استئصاله بالكامل. يندر أن يكون الورم السحائي خبيثاً.

هناك نوعان رئيسيان ل أورام الدماغ و هما: الورم الدماغي الأولي و الورم المنتقل. يبدأ الورم الأولي في الدماغ، أمّا الورم المنتقل ف يبدأ في مكان آخر من الجسم، ثمّ ينتقل إلى الدماغ. هناك نوعان من الأورام الأوليّة و هي: الأورام الحميدة و الأورام الخبيثة. و لا تحوي الأورام الحميدة خلايا سرطانية، أمّا الأورام الخبيثة ف إنّها تحوي خلايا سرطانية. تُسمّى أورام الدماغ الأوليّة الحميدة الأكثر شيوعاً الأورام السحائيّة. و هي تبدأ في غطاء الدماغ الذي يُسمّى الجافية. و هي أكثر شيوعاً لدى النساء منها لدى الرجال. أمّا في المرضى الكبار في السن، ف يجب مراقبة الأورام السحائيّة الصغيرة عند عدم وجود أعراض مهمّة. قد يحتاج الأمر إلى إجراء عملية جراحية ل استئصال الورم السحائي الأكبر حجماً، و من غير المحتمل أن يعود ظهور الورم السحائي عند استئصاله ب الكامل. يندر أن يكون الورم السحائي خبيثاً.

There are two main types of brain tumors: primary and metastatic. Primary tumors start in the brain. Metastatic tumors start somewhere else in the body and move to the brain. There are two kinds of primary tumors: benign and malignant. Benign tumors do not contain cancer cells. Malignant tumors do contain cancer cells. The most common benign primary brain tumors are called "meningiomas." They begin in the covering of the brain called the dura. They are more common in women than in men. In older patients, small meningiomas should be watched if significant symptoms are not occurring. Meningiomas that are bigger or show a tendency to get bigger may need to be removed surgically. If the whole tumor is taken out, it is not likely that the meningioma will come back. Rarely, meningiomas can be malignant.

Figure 4.2 A sample of the un-tokenised data in Arabic, its translation in English, and its tokenised version.

Table 4.1 Data size before and after the tokenising stage.

No	Article	Total Tokens (Un-tokenised Data)	Total Tokens (Tokenised Data)	Difference
1	Adrenal glands cancer	2220	2726	+506
2	Anal Cancer	1765	2198	+433
3	Bone Cancer	1685	2086	+401
4	Brain Cancer	1723	2200	+477
5	Breast Cancer	2289	2777	+488
6	Cervical Cancer	2281	2748	+467
7	Colon Cancer	1500	1922	+422
8	Eye Cancer	2433	3036	+603
9	Gall Bladder Cancer	1673	2081	+408
10	Intestinal Cancer	2014	2511	+497
11	Kidney Cancer	1958	2404	+446
12	Leukaemia	2332	2977	+645
13	Liver Cancer	1976	2482	+506
14	Lung Cancer	1310	1630	+320
15	Nasal Cancer	1719	2155	+436
16	Oesophageal Cancer	2051	2523	+472
17	Oral Cancer	1794	2250	+456
18	Ovarian Cancer	1605	2044	+439
19	Pancreatic Cancer	1580	1908	+328
20	Penile Cancer	932	1169	+237
21	Salivary Gland Cancer	2071	2541	+470
22	Skin Cancer Non-Melanoma	2120	2643	+523
23	Stomach Cancer	1878	2340	+462
24	Testicular Cancer	1379	1775	+396
25	Thyroid Cancer	1974	2444	+470
26	Uterine Cancer	2143	2645	+502
27	Vulvar Cancer	1851	2289	+438
Total		50256	62504	+12248

- **Default:** conjunctions, prepositions, determiners, suffixes and future markers are all individually separated.
- **Default+nodelimit:** conjunctions, prepositions, determiners, suffixes and future markers are all individually separated; no delimiters.
- **Atb:** conjunctions, prepositions, suffixes and future markers are all individually separated (determiners are not segmented).
- **Conj:** only conjunctions are separated (w#, f#).
- **Conj+suff:** conjunctions and suffixes are separated.
- **Prep:** prepositions are separated; any conjunctions are attached to the prep.
- **Group+prep:** prepositions are separated; any conjunctions are attached to the prep but are not delimited.
- **Fut:** the future marker (s+) is separated; any conjunction or prepositions are left attached to s+.
- **Prefix:** All prefixes are separated as one token.
- **Det:** determiners (Al+) are separated; any conjunction or prepositions are # left attached to Al.
- **Suff:** Only suffixes are separated.

Table 4.2 illustrates how the main schemes are applied on the following words: **وبالحسنات** (and+by+their+virtue), **وللبلاد** (and+for+the+countries), **فيمكتبهم** (then+by+librarians+their), **وسنقولها** (and+will+we+say+it).

Table 4.2 The main tokenisation schemes of AMIRA.

The word Scheme	وبالحسنات wbAlHsnAt	وللبلاد wllblAd	فيمكتبهم fbmktbthm	وسنقولها wsnqwlhA
default	w# b# Al# HsnAt	w# l# Al# blAd	f# b# mktbp +hm	w# s# nqwl +hA
default+nodelimit	w b Al HsnAt	w l Al blAd	f b mktbp hm	w s nqwl hA
atb	w# b# AlHsnAt	w# l# AlblAd	f# b# mktbp hm	w# s# nqwl +hA
conj	w# bAlHsnAt	w# llblAd	f# b mktbthm	w# snqwlhA
conj+suff	w# bAlHsnAt	w# llblAd	f# b mktbp +hm	w# snqwl +hA
Prep	w#b# AlHsnAt	w#l# AlblAd	f#b# mktbthm	wsnqwlhA
group+prep	wb# AlHsnAt	wl# AlblAd	fb# mktbthm	wsnqwlhA
fut	wbAlHsnAt	wllblAd	fbmktbthm	w#s# nqwlhA
prefix	w#b#Al# HsnAt	w#l#Al# blAd	f#b# mktbthm	w#s# nqwlhA
det	w#b#Al# HsnAt	w#l#Al# blAd	fbmktbthm	wsnqwlhA
suff	wbAlHsnAt	wllblAd	fbmktbp +hm	wsnqwl +hA

AMIRA is applied to 27 cancer articles where all prefixes, such as conjunctions, future markers, and prepositions, are segmented off the lexical item. The AI determiners and suffixes are not tokenised because this increases the ambiguity and sparsity of the text, as there are more than 127 suffixes in Arabic (Sawalha and Atwell, 2009). Figure 4.3 displays a sample of the tokenisation result, where errors are highlighted in grey.

Uterine cancer is cancer that begins in the uterus. This program will focus on the most common type of uterine cancer, which is endometrial cancer. Endometrial cancer begins in the lining of the uterus. Cancerous cells spread to different parts of the body through blood vessels and lymph channels. It is usually impossible to specify the .cause of cancer in an individual patient	سرطان الرحم هو السرطان الذي يبدأ في الرحم. يهتم هذا البرنامج بالنوع الأكثر شيوعاً من سرطان الرحم، وهو السرطان البطانية الرحمية. تبدأ السرطانة البطانية الرحمية في بطانة الرحم الداخلية. تنتشر الخلايا السرطانية إلى أجزاء مختلفة من الجسم عن طريق الأوعية الدموية والقنوات اللمفية. ويكون من المستحيل تحديد السبب الدقيق للإصابة بالسرطان لدى مريض ب عينة عادة
--	--

Figure 4.3 A sample of the tokenisation task result.

In the above example, AMIRA has missed tokenising lexical items which start with the preposition ب (b) such as بالنوع (bAlnwE – type) and بالسرطان (bAlsrTAn – by cancer) and lexical item which starts with the conjunction و (w) such as وهو (whw – and it). However, AMIRA has tokenised incorrectly the lexical item اللمفية (Allmfyp – lymphatic) by adding the letter ا (A) after the determiner ال (Al).

The results of AMIRA's tokenisation of our corpus are evaluated in terms of three measures: precision, recall, and F-measure, using the following equations:

$$Precision = \frac{\text{the number of words that have been tokenised correctly}}{\text{the number of words that have been tokenised}} \quad \text{Equation 4.1}$$

$$Recall = \frac{\text{the number of words that have been tokenised correctly}}{\text{the number of words that need to be tokenised}} \quad \text{Equation 4.2}$$

$$F - \text{measure} = \frac{2 * recall * precision}{recall + precision} \quad \text{Equation 4.3}$$

The AMIRA tool has achieved 91.30%, 88.53%, and 89.90% for precision, recall, and F-measure, respectively. Two categories of errors are identified: false positive errors and false negative errors,

which are described below. False positive errors occur when AMIRA tokenises a lexical item that does not need to be tokenised. Five different false positive errors have been identified and described below.

- Lexical items starting with the letter *l*

AMIRA incorrectly tokenises some lexical items which start with the letter *l*, confusing it with the determiner *al*; in these lexical items the letter *l* is part of the lexical item. Examples of these lexical items are لقاح (vaccine), ليست (not), لعابية (salivary), ليزرية (laser), and لمفية (lymphocytes); these lexical items are tokenised into two terms: ل يزرية, ل عابية, ل يست, ل قاح, and ل مفية.

- Lexical items starting with the letter *f*

As in Arabic the letter *f* (ف) can also refer to a preposition ‘in’, AMIRA is not able to distinguish whether the letter *f* refers to a preposition or is part of the lexical item. Examples of these lexical items are فحص (check), فعالية (effectiveness), فلتر (filter), فالوب (fallopian), فقدان (loss), فص, (lobe), فلورسيني and (Florenzi) which are tokenised into ف قحان, ف الو ب, ف لتر, ف عالية, ف حص, and ف لورسيني.

- Lexical items starting with the letter *w*

Similarly, the letter *w* (و) can be a conjunction meaning ‘and’ as in the following lexical items: وهذا (and this) and السرطان (and the cancer) It can also be part of the lexical item as in these examples: وراثية (genetic), ورم (tumour), وظائف (functions), وجع (pain), وريدي (vascular), وجود (existence), and وعاءان (vessels). These are incorrectly tokenised into two lexical items: وراثية, ورم, وظائف, وجع, وريدي, وجود, وعاءان. This problem also occurs with foreign items such as ويلمز (Wilms) and ويدمان (Wiedemann).

- Lexical items starting with the letter *l* after the *Al* determiner

One of the most common false positive errors occurs when tokenising lexical items where the first letter after the ال (*Al*) determiner is ل (*L*). Examples of these lexical items are اللعابية (AlIEAby – salivary), اللمفية (AlImfyp – lymphatic), اللوزتين (AlIwztyn – tonsils), and اللوكيميا (AlIwkymyA – leukaemia). Some of these errors may be related to the limited data set used by AMIRA’s classifier. AMIRA adds the letter ا (*A*) after the determiner in these lexical items, so the incorrect results of tokenising these lexical items are الالعبية (AlAlIEAby), الاللمفية (AlAlImfyp), الالوزتين (AlAlIwztyn), and الالوكيميا (AlAlIwkymyA). These errors were corrected manually before moving to the next task.

- Lexical items starting with the letter *k*

AMIRA has only a few errors when it comes to the letter *k*. The letter *k* can be a preposition as in the term: كسرطان (such as cancer) or it can be an original part of the word as in the words كبِد (liver), كتل (masses), and كروموسومات (chromosomes). Therefore, the inaccurate results of tokenising these lexical items are ك تل, ك بد, and كروموسومات.

- Lexical items that start with the letter *b*

AMIRA also displays a few errors in tokenising lexical items that start with the letter ب (*b*) which can be a preposition such as in the word: بسرطان (by cancer) or it can be an original letter of the lexical item. For instance, AMIRA split the letter *b* which is an original letter in lexical items like بلغم (phlegm), بوليب (polyp), and بكويث (Beckwith).

False negative errors occur when AMIRA misses tokenising a lexical item that needs to be tokenised. Five different false positive errors are identified and listed below.

- Lexical items that start with the preposition *b*

The most common false negative errors related to cases where lexical items that start with the preposition *b* are not segmented off. Examples of these lexical items are بالسرطان (bAlsrTAn – by cancer), بالإضافة (bAlAdAfp – in addition), بالدهون (bAldhwn – with fats), and باليود (bAlywd – with iodine).

- Lexical items that start with the preposition *k*

AMIRA misses tokenising some lexical items that start with the preposition *k*. Examples of these lexical items are كالأنف (like nose), كالسرطان (like cancer), and كالإسهال (like diarrhoea).

- Lexical items that start with the preposition *l*

AMIRA misses tokenising some lexical items that start with the preposition *l*. Examples of these lexical items are لاكتشاف (to discover), لاستئصال (to eradicate), and لائل (to liquid).

- Lexical item that start with the conjunction *f*

AMIRA misses tokenising many lexical items that start with the *f* conjunction. Examples of these lexical items are فالسرطان (and cancer), فالمرأة (and women), فيبدأ (and start), and فالناس (and people).

- Lexical items that start with the conjunction *w*.

AMIRA misses tokenising some lexical items that start with the conjunction *w*, for example, the lexical items *واللسان* (and the tongue), *وقاع* (and the bottom of), and *وغدة* (and the gland).

A simple technique that would boost the performance of the AMIRA tokeniser and minimise the previous errors is matching the rest of each of the lexical items after deleting the prefix against an Arabic dictionary. If a match is found, then the deleted part of the lexical item is a prefix. Although this simple technique can minimise the rate of errors, it fails to work in some cases where the rest of the lexical item is a correct Arabic lexical item. For instance, the remainders after tokenising the letter *l* in lexical items like *لديك* (you have), and *لمس* (touch) and the letter *f* in *فيصل* (and arrive) are correct Arabic lexical items. In such cases, AMIRA requires knowledge of the contextual domain; it is not possible even for a human to determine whether these lexical items need to be tokenised without understanding the context. Another approach is to develop a machine learning algorithm using a very large set of data to improve tokenisation. To address the false negative errors related to the preposition *ب* (*b*) followed by the determiner *ال* (*Al*), one can create a gazetteer of those lexical items as their occurrences is limited in the Arabic. A study of ANERcorp corpus, which consists of around 150,000 tokens (Benajiba *et al.*, 2007) produced 1,104 lexical items start with *بالـ* (*bAl*). and in only 21 of these, *بالـ* (*bAl*) is part of the original lexical item, and nine of these 21 are non-Arabic lexical items. The rest of the 21 lexical items are a repetition of only four Arabic lexical items, which are *بالغة* (*bAlghp* – exaggerate), *بالغ* (*bAlgh* – adult), *بال* (*bAl* – shabby) and *بالي* (*bAly* – shabby).

4.4.1.2 Part of speech tagging

Two different tools are used to label each token in the corpus: AMIRA and MADAMIRA. AMIRA was used to test its performance and applied on a set of 5,119 tokens. Subsequently, a more recent tool, MADAMIRA, has been made available to researchers and applied to a larger set consisting of 62,504 tokens (Table 4.1). The finding of both tools are described below.

AMIRA is a shallow syntactic parser toolkit consisting of a tokeniser, POS tagger, and base phrase chunker. It is based on supervised learning and relies on surface data to learn generalisations. The parser includes three different tag sets: the Bies tagset, extended reduced tagset (ERTS), and extended reduced tagset and person information (ERTS_PER). The Bies tagset, which was developed by Ann Bies and Dan Bikel, consists of 24 tags (Diab, 2009). It ignores certain Arabic distinctions; for example, it treats the dual form, a common form in the Arabic language, as a plural. It also cannot specify gender in both verbs and nouns. The ERTS tagset, which has 72 tags and provides additional

morphological features to the Bies tagset, can handle number (singular/dual/plural), gender (feminine/masculine), and definiteness (i.e., the existence of the definite article or not). In addition to the tags in the ERTS tagset, the ERTS_PER specifies the use of the first, second, and third person voice.

The ERTS, which is selected for the POS tagging task, has many relevant morphological features for our corpus, while person information is a less important feature, as our corpus is limited to the third person voice. A sample is shown in Figure 4.4. In this example, AMIRA assigns a noun tag to the two adverbs: **خلف** (behind) and **أمام** (in front of). It also assigns an adjective tag to the genitive noun **المعدة** (stomach) and fails in assigning the plural noun tag NNS to the lexical item **عوامل** (factors). The POS tagger of AMIRA, which is estimated using the formula below, achieves an accuracy of 84.09% (see Equation 4.4). However, Arabic POS taggers still need further research to at least reach the 97.3% accuracy of the Stanford POS tagger for the English language, (Manning, 2011)

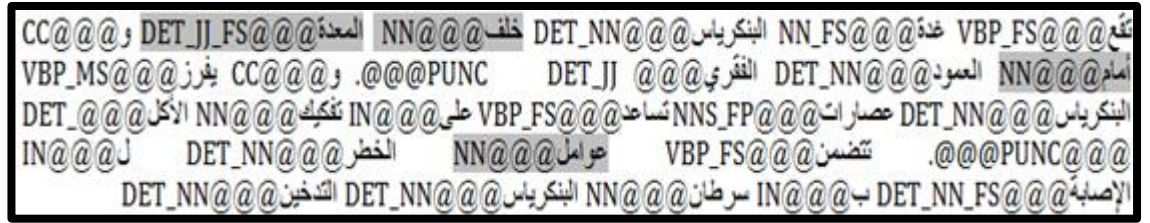


Figure 4.4 A sample of the POS tagging task result.

$$Accuracy = \frac{\text{the number of correctly tagged tokens}}{\text{the total number of tokens}} \quad \text{Equation 4.4}$$

Due to the complex and challenging nature of the Arabic language, AMIRA performs less favourably than English parsers in the following areas:

- Broken plurals

Arabic has three types of plurals: the broken plural, the sound masculine plural, and the sound feminine plural. The most common type is the broken (irregular) plural, constituting about half of all plurals in Arabic (Habash, 2010). Furthermore, AMIRA has limited capability to assign an appropriate POS tag to broken plurals, as 32.02% of AMIRA errors among our corpus are related to broken-plural lexical items. For instance, AMIRA assigns a singular feminine lexical item tag (DET_NN_FS) to the broken-

plural lexical items ‘الأوعية’ ‘utensils’, ‘الأنسجة’ ‘tissues’, and ‘الأكنية’ ‘ducts’. It also failed to assign a plural noun tag (NNS) to most of the other broken-plural lexical items. Examples of these lexical items are ‘الأطباء’ ‘doctors’, ‘سُبُل’ ‘ways’, and ‘خلايا’ ‘cells’. Broken plurals can be formed using more than 20 morphological patterns. Furthermore, an Arabic lexical item might have more than one plural. For instance, the lexical item ‘أسد’ ‘lion’ has five different broken-plural forms (أسود – أسود – أسد – أسدة – أسد). Therefore, it can be quite difficult to identify a solution for broken-plural POS tagging. To improve the performance of broken-plural POS tagging machine-learning classifier techniques, such as neural networks or a decision tree are employed. In the literature, Goweder *et al.* (2004) examined different methods and they concluded that the dictionary and decision tree methods achieved the highest results in identifying broken plurals.

- Adverbs

In Arabic, there are two main types of adverbs: those describing time and others referring to location. AMIRA assigned a noun tag (NN) to most adverbs in our corpus. Examples of these adverbs are ‘خلف’ ‘behind’, ‘أسفل’ ‘at the bottom of’, and ‘بعد’ ‘after’. We propose, as future work, to create an adverb gazetteer and use it as a binary feature to feed the machine-learning classifier.

- Adjective and genitive nouns

One of the most frequent errors in AMIRA’s POS output is assigning an adjective tag (JJ) to genitive nouns (المضاف إليه). For instance, AMIRA assigns a JJ tag to the word ‘stomach’ in the phrase ‘سرطان المعدة’ ‘cancer of the stomach’, the word ‘patient’ in the phrase ‘فرصة المريض’ ‘the patient’s chance’, and the word ‘appetite’ in the phrase ‘نقصان الشهية’ ‘loss of appetite’. There are some grammatical differences between adjectives and genitive nouns in Arabic grammar. Adjectives and the nouns they modify must agree in number (singular/dual/plural), mood (indicative/subjunctive/genitive), and indefiniteness or definiteness (presence of the definite article). In the above examples, the adjectives and the nouns they modify disagree in both mood and indefiniteness or definiteness. Using these grammatical differences as features in the data training phase will improve the task of differentiation between adjectives and genitive nouns. Because of these limitations of AMIRA, an alternative tool is investigated to run the second experiment. MADAMIRA is a tool for morphological analysis and disambiguation of the Arabic language (Pasha *et al.*, 2014). It is a combination and refinement of two previous and common tools used for Arabic processing, MADA (Habash *et al.*, 2009) and AMIRA (Diab *et al.*, 2007). The POS tagset of MADAMIRA contains 35 different tags: one tag for nouns, two tags for numbers, one tag for proper nouns, three tags for adjectives, three tags for adverbs, five tags for pronouns, two tags for verbs, 10 tags for particles, one tag for prepositions, one tag for abbreviations, one tag for punctuation, two tags for conjunctions, one tag for interjections, one tag for

digital numbers, and one tag for foreign/Latin words. Table 4.3 lists the tags used in MADAMIRA and their definitions. Unlike the first experiment, the output of the POS tagging task is not corrected manually, as MADAMIRA achieved an accuracy of 96% when it was applied to a blind test data set consisting of 25,000 words for modern standard Arabic (Pasha *et al.*, 2013). We believe that a 4% error rate cannot significantly affect the results of our classifier, especially as the POS tags comprise only one feature among more than six features that can be used to train our classifier. Moreover, POS taggers tend to make systematic errors, and systematic errors cannot significantly affect the performance of the classifier, as the same errors appeared in the training data.

4.4.1.3 Data Annotation

Data annotation is a crucial step in any supervised learning-based NER system. It is a laborious and time-consuming process that requires human effort to carry out the task of annotation. In the context of NER, data annotation is the process of labelling each token in the data with an appropriate tag. The tags should be listed and predefined before starting the data annotation task. Our tokenised corpus, which included 62,500 tokens, has been manually annotated by the researcher to maintain coherence. During the data annotation, four NEs are recognised as relevant to the cancer domain. These are disease name (D), symptoms (S), treatment methods (T), and diagnosis methods (G). In the NER literature, the most commonly 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 lexical item as being inside the NE, and the tag O marks the lexical item as being outside the NE, while in the IOB tagging scheme the extra B tag marks the beginning of the NE. Table 4.4 illustrates an example of how the sentence, ‘Salivary gland cancer is rare’ is tagged by IO and IOB schemes.

The IO tagging scheme is used in the data annotation step. Although this scheme cannot determine the boundary in the case of two NEs from the same class appearing next to each other, this does not affect the performance of this step, as no such two NEs from the same class appear next to each other in our dataset. Furthermore, IO outperforms the IOB scheme in terms of cost and running 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 NE classes. Our corpus contains 1,279 disease entities with total of 2,797 tokens; each disease entity is represented on average by around two tokens, while the disease entity constitutes around 4.47% of the whole corpus (Table 4.5). The number of entities representing the symptoms is limited to 333 entities out of 1,264 tokens, and each symptom entity is expressed on average by almost four tokens representing just over 2% of the whole corpus. The numbers of the treatment and diagnosis method entities are 693 and 501 with total of 1,309 and 1,147 tokens, respectively.

Table 4.3 List of MADAMIRA POS Tags.

No	POS Tag	POS Tag Definition
1	Noun	noun
2	noun_num	Number
3	noun_quant	
4	noun_prop	Proper Nouns
5	adj	Adjectives
6	adj_comp	
7	adj_num	
8	adv	Adverbs
9	adv_interrog	
10	adv_rel	
11	pron	Pronouns
12	pron_dem	
13	pron_exclam	
14	pron_interrog	
15	pron_rel	
16	verb	Verbs
17	verb_pseudo	
18	part	Particles
19	part_dem	
20	part_det	
21	part_focus	
22	part_fut	
23	part_interrog	
24	part_neg	
25	part_restrict	
26	part_verb	
27	part_voc	
28	prep	Prepositions
29	abbrev	Abbreviations
30	punc	Punctuation
31	conj	Conjunctions
32	conj_sub	
33	Interj	Interjections
34	digit	Digital Numbers
35	latin	Foreign/Latin

Table 4.4 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 treatments and diagnosis methods are represented on average by less than two (1.88) and just over two (2.28) entities, respectively, and they constitute 2.09% and 1.83% of the data, respectively. The disease entity category gained the highest number of entities; this is because disease is the most frequent entity type used in our corpus. However, the symptom entity is often expressed in terms of more than four tokens and up to 15 tokens as in the following example: ‘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’. A summary of the total number of entities and tokens extracted from the corpus for each article is given in the tables below.

Table 4.5 Total number of entities and tokens in our corpus.

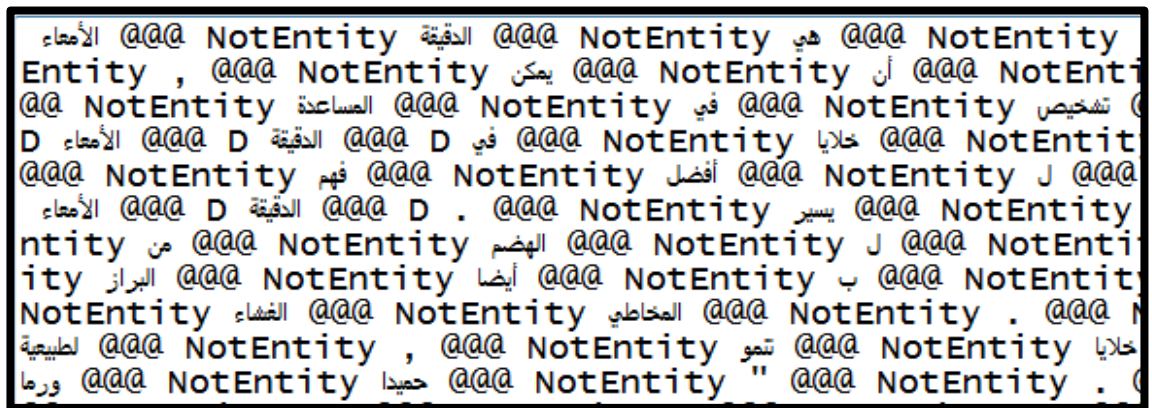
Named Entity Class	Number of Entities	Number of Tokens	Ratio of Tokens to Data
Disease Names	1279	2797	4.47%
Symptoms	333	1264	2.02%
Treatment Methods	693	1309	2.09%
Diagnosis Methods	501	1147	1.83%
Total	2806	6517	10.41%

To ensure coherence in the data annotation stage, a set of annotation guidelines is applied and listed below:

- Different types of a cancer disease are tagged as disease entities. For example, acute lymphocytic leukaemia, acute myeloid leukaemia, chronic lymphocytic leukaemia, and chronic myeloid leukaemia are types of leukaemia and therefore tagged as leukaemia.
- Related names of a given disease are tagged as a disease entity. For example, blood cancer and leukaemia, both are tagged as disease.
- Collective nouns of a disease, such as blood cancers and brain cancers, are tagged as a disease.

- General lexical terms (e.g. cancer) that do not refer to a specific type of entity are not tagged as a disease.
- The symptoms are tagged as symptom entities regardless of the number of tokens used to express them; for example, symptoms such as a lump in the area of the ear, cheek, jaw, lip, or inside the mouth are tagged as a symptom.
- The anaphoric references are not included and hence are not annotated.

Figure 4.5 shows a sample of the annotated data. As shown, each token in the data is given an appropriate tag, while the symbol '@@@' is used for programming purposes.



@@@ NotEntity الدقبة @@@ NotEntity هي @@@ NotEntity
Entity , @@@ NotEntity يمكن @@@ NotEntity أن @@@ NotEntity
@@@ NotEntity المساعدة @@@ NotEntity في @@@ NotEntity تشخيص
D @@@ D الدقبة @@@ D في @@@ NotEntity خلايا @@@ NotEntity
@@@ NotEntity فهم @@@ NotEntity أفضل @@@ NotEntity ج @@@
@@@ D الدقبة @@@ D . @@@ NotEntity يسير @@@ NotEntity
ntity من @@@ NotEntity الهضم @@@ NotEntity ج @@@ NotEntity
ity @@@ NotEntity أيضا @@@ NotEntity ب @@@ NotEntity
NotEntity الغشاء @@@ NotEntity المخاطي @@@ NotEntity . @@@
خلايا @@@ NotEntity , @@@ NotEntity تنمو @@@ NotEntity
ورما @@@ NotEntity حميدا @@@ NotEntity " @@@ NotEntity .

Figure 4.5 A sample of the annotated data.

Table 4.6 Disease entities and tokens in the corpus.

Article	Extracted disease entities*	Total number of tokens	Article	Extracted disease entities*	Total number of tokens
Adrenal gland cancer	40	86	Lung Cancer	32	74
Anal Cancer	46	92	Nasal Cancer	29	75
Bone Cancer	54	108	Oral Cancer	59	119
Brain Cancer	62	140	Ovarian Cancer	48	97
Breast Cancer	44	88	Pancreatic Cancer	34	69
Cervical Cancer	46	138	Penile Cancer	32	69
Colon Cancer	37	74	Salivary Gland Cancer	41	125
Esophageal Cancer	48	96	Skin Cancer	71	157
Eye Cancer	68	164	Stomach Cancer	35	90
Gall Bladder Cancer	40	70	Testicular Cancer	25	50
Intestinal Cancer	45	128	Thyroid Cancer	59	131
Kidney Cancer	48	98	Uterine Cancer	62	106
Leukaemia	86	176	Vulvar Cancer	39	79
Liver Cancer	49	98	Total	1279	2797

* Some may be repeated entities and anaphoric references are not included.

One entity may consist of more than one token.

Table 4.7 Number of symptoms entities and tokens extracted.

Article	Extracted symptom entities*	Total number of tokens	Article	Extracted symptom entities*	Total number of tokens
Adrenal gland cancer	2	9	Lung Cancer	25	76
Anal Cancer	5	31	Nasal Cancer	18	82
Bone Cancer	4	8	Oral Cancer	16	68
Brain Cancer	19	67	Ovarian Cancer	16	47
Breast Cancer	8	38	Pancreatic Cancer	7	28
Cervical Cancer	11	43	Penile Cancer	8	47
Colon Cancer	13	35	Salivary Gland Cancer	6	56
Esophageal Cancer	11	38	Skin Cancer	3	7
Eye Cancer	12	40	Stomach Cancer	15	46
Gall Bladder Cancer	19	56	Testicular Cancer	16	93
Intestinal Cancer	19	42	Thyroid Cancer	12	57
Kidney Cancer	11	46	Uterine Cancer	8	25
Leukaemia	19	57	Vulvar Cancer	8	43
Liver Cancer	22	79	Total	333	1264

* Some may be repeated entities and anaphoric references are not included.

One entity may consist of more than one token.

Table 4.8 Number of treatment entities and tokens extracted.

Article	Extracted treatment method entities*	Total number of tokens	Article	Extracted treatment method entities*	Total number of tokens
Adrenal gland cancer	22	41	Lung Cancer	6	9
Anal Cancer	24	46	Nasal Cancer	17	29
Bone Cancer	28	48	Oral Cancer	21	40
Brain Cancer	33	57	Ovarian Cancer	16	36
Breast Cancer	50	117	Pancreatic Cancer	24	40
Cervical Cancer	17	33	Penile Cancer	9	20
Colon Cancer	17	25	Salivary Gland Cancer	23	41
Esophageal Cancer	36	64	Skin Cancer	24	44
Eye Cancer	34	65	Stomach Cancer	36	58
Gall Bladder Cancer	24	44	Testicular Cancer	23	44
Intestinal Cancer	21	36	Thyroid Cancer	27	51
Kidney Cancer	29	63	Uterine Cancer	30	55
Leukaemia	38	75	Vulvar Cancer	29	47
Liver Cancer	35	81	Total	693	1309

* Some may be repeated entities and anaphoric references are not included.

One entity may consist of more than one token.

Table 4.9 Number of diagnostic entities and tokens extracted.

Article	Extracted diagnosis method entities*	Total number of tokens	Article	Extracted diagnosis method entities*	Total number of tokens
Adrenal gland cancer	26	62	Lung Cancer	16	31
Anal Cancer	22	48	Nasal Cancer	22	51
Bone Cancer	15	34	Oral Cancer	18	43
Brain Cancer	13	26	Ovarian Cancer	9	20
Breast Cancer	15	30	Pancreatic Cancer	16	26
Cervical Cancer	20	47	Penile Cancer	4	18
Colon Cancer	28	56	Salivary Gland Cancer	23	58
Esophageal Cancer	27	60	Skin Cancer	12	17
Eye Cancer	25	69	Stomach Cancer	24	51
Gall Bladder Cancer	22	50	Testicular Cancer	9	19
Intestinal Cancer	26	56	Thyroid Cancer	22	51
Kidney Cancer	18	42	Uterine Cancer	21	46
Leukaemia	7	25	Vulvar Cancer	21	60
Liver Cancer	20	51	Total	501	1147

* Some may be repeated entities and anaphoric references are not included.

One entity may consist of more than one token.

4.4.2 Data Analysis step

The data analysis step includes three main tasks: frequency analysis, collocation, and concordance. Frequency analysis, an important part of task when exploring a corpus, often includes the use and analysis of lexical items to illustrate and organise their frequency. Sinclair (1991, p.30) explained that “anyone studying a text is likely to need to know how often each different lexical item form occurs in it”. Collocation analysis, another important technique is used to study words which appear together and convey meaning by association. As Firth (1957) explains, you know a word through the company it keeps. Nowhere is this more evident than when analysing the use of collocations in a language. In many ways, collocates come so naturally that they are predictable (Crystal, 2004). However, this does not justify ignoring collocation analysis. If anything, the natural occurrence and predictability of certain collocations over others makes collocation analysis a highly significant step in analysing and understanding a text as a whole. Concordance analysis focuses on analysing the concordance lines. Bennet stated that “[c]oncordance lines are all the instances of a word or phrase in the corpus” (2010, p.17). Sorted concordance lines allow the analyser to “see more extended context of the word as it appears in the corpus” (Bennet, 2010, p.17).

4.4.2.1 Frequency analysis

Frequency analysis is used to study our corpus and to indicate the most frequent tokens in the data. A token is an individual occurrence of a linguistic unit in an oral or written text. This differs from type, as type is “the number of distinct words”, whereas a token is ‘the total count of running words’ (Fry, 2011, p.4). As defined by the Oxford Dictionary, frequency is the ‘rate at which something occurs over a particular period of time or in a given sample’. Therefore, token frequency refers to how often lexical items/words are repeated in a text, as opposed to type, which just illustrates how many different lexical items the text contains. In this study, a distinct token is a single lexical item including suffixes and excluding prefixes, given that text under study has been tokenised.

Frequency is an important aspect of textual analysis because information about word frequency helps us identify key characteristics of a given text or form of communication. Harvey (2013, p.56) explained that for “the analyst interested in examining discursial patterns and commonalities, the value of a wordlist reside [sic] in its ability to provide evidence of the “markedness” of particular discourses or attitudes”. Frequency can illustrate language biases, common concerns, or personal inhibitions, for example. Therefore, frequency reveals the language users’ preferences for specific stances, attitudes, and opinions. This helps the analyser to better understand the personal and cultural context of a text and its value within that context. The result is that the analyser has a better understanding of how the language user makes sense of the world around them. By highlighting commonly used words,

frequency analysis also draws attention to those words that are important to people in a particular field, culture, or context. A word preference may be due to interest or lack of interest in a subject, not due to personal biases and opinions. For example, Baker (2010) found that the word ‘homosexuality’ is more common than the word ‘heterosexuality’, as it occurs over twice as often in the British National Corpus. This could be attributed to simple bias and pejorative usage, as Baker notes. He added that it is also probably due in part to the fact that heterosexuality is considered normal and is therefore not a debated topic, whereas homosexuality is considered unusual and is therefore discussed more frequently (Baker, 2010). This leaves room for context to alter word frequency. For example, the term ‘leukaemia’ is not in common usage. A scan of the British National Corpus revealed 443 incidents in 100 million words or a frequency of five per million. Meanwhile, in our corpus, the frequency is 16 incidents out of 62 thousand lexical items or a frequency of 258 per million. This is not only because the average person does not need to discuss leukaemia often but also because people working in medical fields need to discuss it more, as it is still a poorly understood form of cancer. This shows the bias comes from both sides; average people discuss it less, but people in medical fields discuss it more frequently. However, frequency-sorted wordlists, while invaluable tools for corpus analysis, are still just tools. To make use of them, we need to be selective about the words that are analysed and studied, hence the concept of keywords. These were defined by Harvey (2013, p. 57-58) as “word forms that occur in one particular corpus with a greater significant frequency than in another dataset”. These words can often be considered important to the meaning of the text and their frequency defines its theme. For example, it is easy to see from observing the keywords in my analysis that the analysed text is themed around health, disease, illness, and cures, reflecting the medical context of the corpus. The main benefit of keywords, as identified by Seale and Charteris-Black (2010), is their ability to highlight aspects of text that may be difficult to see when reading the text casually.

This study uses word lists, referred to as lexical items, to analyse token frequency in our corpus. A frequency list provides a record of how often each individual lexical item occurs in the analysed text; it can be organised in three different ways: order of first occurrence, alphabetical order, or order of frequency. This study employs frequency-ordered lists, showing which lexical items occur more often. This is because alphabetical order is “built mainly for indexing purposes” and first occurrence order is “a quick guide to the distribution of words in a text” (Baron *et al.*, 2009, p.1), whereas frequency-ordered lists draw attention to the most common lexical items, which is the focus of our study. This frequency analysis is carried out to reveal the most common lexical items in a text. The most frequent keywords are used in a collocation analysis to define their situational context and carried out by the concordance analysis tool, aConCorde 0.4.3, which is a multilingual concordance tool originally designed for Arabic concordance analysis (Roberts *et al.*, 2005).

Our corpus consists of 4,870 distinct and unique lexical items. As shown in Table 4.10, only one lexical item is repeated more than 2,000 times and only two lexical items are repeated between 1,500 and 1,999 times, while three and six lexical items are repeated between 1,000 to 1,499 times and 500 to 999 times, respectively. As we see in Table 4.10, the relationship between the frequency of these lexical items and the number of lexical items are repeated in an inverse relationship; whenever the frequency increases, the number of lexical items decreases. For example, the lexical items that appear only once in our corpus constitute around 43.34% of our corpus size, with a total of 2,111 lexical items out of 4,870 lexical items.

Table 4.10 The frequency distribution of the lexical items in the corpus.

Frequency	Number of lexical items
more than 2000	1
between 1500-1999	2
between 1000-1499	3
between 500-999	6
between 300-499	13
between 200-299	9
between 100-199	45
between 50-99	89
between 30-49	133
between 20-29	157
between 10-19	375
between 5-9	537
between 3-4	630
twice	759
once	2111
Total	4870

Table 4.11 lists the 30 most frequent lexical items in our corpus. Among those lexical items, the most frequent lexical item is the conjunction ‘and’, which appeared 2,837 times in our corpus. Usually, the highest frequency of lexical items in any corpus are stop words, which provide less meaning. However, some of the 30 most frequent words in Table 4.11 are more informative (highlighted in grey shading). The frequency analysis has yielded the creation of different lists, namely a list of stop words, gazetteers, and a list of lexical markers. The stop word list contains tokens that appear in any open-domain corpus and impart very little meaning on their own, namely prepositions, articles, and conjunctions. Gazetteers are dictionaries that collect lists of relevant NEs to the corpus. Four different gazetteers were created for disease names, symptoms, treatment methods, and diagnostic methods.

Lexical markers are the lexical tokens that can indicate the presence of the NEs. For example, the lexical items ‘الشعور’ and ‘ألم’ are lexical markers for the symptom entity.

Table 4.11 The 30 most frequent lexical items in our corpus.

Token	Translation	Frequency	Token	Translation	Frequency
و	and	2837	من	from	1941
ب	by, with, in	1613	في	In	1290
ل	to , for , so	1143	سرطان	cancer	1019
أو	or	807	السرطان	the cancer	765
على	on	757	إلى	to	735
أن	that	636	المعالجة	the treatment	500
يمكن	could	457	الجسم	the body	452
قد	may	437	الخلايا	the cells	354
إذا	if	344	ف	and , so	339
هذه	this (feminine)	339	إن	If	326
هذا	this (masculine)	322	التي	which(feminine)	315
عن	about, on	313	الورم	the tumour	308
هو	he , this	303	ما	what	293
أيضا	also	280	الطبيب	the doctor	257
المريض	the patient	238	هي	she , this	233

4.4.2.2 Collocation Analysis

Collocation is a crucial tool of textual analysis. Although keywords provide various methodological and analytical advantages, they are not in and of themselves an analysis of the corpus as a whole. They are, again, additional instruments towards the effort of analysis. Keywords indicate the most promising lexical items and expressions in a text, allowing readers to identify themes, notable concepts, and the most discussed topics. However, to interpret the text fully, we need to analyse the keywords in their context. In order to do this, we employ collocation analysis. Collocation focuses on co-occurrence of keywords with other expressions and concepts, revealing which lexical items are associated with each other and how this alters the meaning of the keyword. It provides “a way of understanding meanings and associations between words which are otherwise difficult to ascertain from a small-scale analysis of a single text” (Baker, 2006, p. 96). This additional layer of meaning may illustrate how an individual or culture views the keyword, at least within the context of the text’s theme.

Gledhill (2000) stated that there is no single definition of collocation that covers all its meanings within the field of linguistics. Therefore, the statistical and textual meaning of collocations will vary. Van Roey (1990) summarised collocation as a linguistic phenomenon that demonstrates not which lexical

items necessarily belong together for grammatical or conceptual purposes but which lexical items are connected by people's preference of usage. This creates an individual or cultural context for collocates, rather than a definitional or grammatical context. This context will shift from text to text, based on the text's theme and author. For this reason, many researchers observe collocation alongside co-occurrence and statistical probability. Collocation is, in the context of statistical textual analysis, meaningful terms that co-occur within a certain distance of a keyword, as established by the analyser. If a co-occurrence fits within the parameters of statistical probability, it can therefore be considered accidental or unrelated, whereas if the co-occurrence exceeds statistical probability, it is deemed meaningful and therefore listed as a collocation. In short, collocation is a combination of co-occurrence and recurrence of meaningful lexical items with keywords (Gledhill, 2000). Collocations also develop within a text. Reformulation, repetition, and paraphrases of synonymous expressions within a text can create various expressions that illustrate the same theme. They may have identical context or the same textual triggers, also related by distance in the text, as noted by Hoey (1991).

Collocations have various uses, as noted by Manning and Schütze (1999). They can be applied to natural language generation to confirm that the expressions created sound perfectly natural by preventing uncommon collocates and employing common ones. They can be applied to computational lexicography in order to automatically identify and list collocations for dictionaries. They can be applied to parsing to give preference to parses that have natural collocations. They are also highly important to corpus linguistics research in identifying the linguistic habits of languages, regions, individuals, research fields, etc. Beyond that, they have also been found to have applications in lexical item sense disambiguation (Ide and Véronis, 1998), language teaching (Nesselhauf, 2003), and machine translation (Smadja *et al.*, 1996). Its extensive variety of uses highlights the significance of collocation analysis in textual analysis as well as the significance of collocations in language as a whole.

The text2ngram tool which is a set of tools for extracting N-grams from raw corpus up to 255-grams is used to perform the collocation analysis. Table 4.12 demonstrates the distribution of the collocations and their frequencies among our corpus. 5,559 collocations are identified in our corpus. Of those, four collocations occur more than 200 times in our corpus and 2,345 collocations appeared only twice among our data. As we observe in Table 4.12, the relationship between the number of collocations and their frequency is an inverse relationship where, whenever the frequency increases, the number of collocations decrease. For example, the collocations that appeared only twice in our corpus constitute around 42.18% of all the collocations in our corpus.

Table 4.13 lists the ten most recurrent collocations in our data. The most frequent collocation in our data is the phrase “يَمَكُنْ أُنْ” (maybe), which appeared 283 times. Some of the highest frequency

collocations in our corpus are less informative than others. For instance, the most repeated collocation is the term “يمكن أن” (maybe). This collocation does not indicate a specific named entity of interest. On the other hand, the collocation “الإصابة بـ” (infection with) usually indicates that the following lexical item is a disease entity. The most informative collocations are shaded with grey in Table 4.13

Table 4.12 The frequency distribution of the collocations among our corpus.

Frequency	Number of collocations
More than 200	4
Between 100–199	9
Between 50–99	49
Between 30–49	99
Between 10–29	476
Between 5–9	1016
3 and 4 times	1561
Twice	2345
Total	5559

Table 4.13 The ten most recurrent collocations in our data.

Collocation	Translation	Frequency	Collocation	Translation	Frequency
يمكن أن	maybe	283	ب سرطان	by cancer	242
من أجل	for the sake of	217	الإصابة بـ	infection with	209
إذا كان	if it was	170	ب اسم	By the name of	134
من الجسم	of body	127	و هو	And it	125
و قد	and may	125	الخلايا السرطانية	cancer cells	125

Table 4.14 lists the most 20 informative collocations in terms of their frequency among our corpus. As the table demonstrates, these collocations are related to NE of specific interest. For instance, the collocations ‘ب سرطان’ ‘by cancer’, ‘الإصابة بـ’ ‘infection with’, and ‘معالجة سرطان’ ‘curing the cancer of’ are indicators of the disease entities, while the collocations ‘ابيضاض الدم’ ‘leukaemia’, ‘سرطان الدرق’ ‘thyroid cancer’, and ‘سرطان الفم’ ‘oral cancer’ are disease entities. Regarding the remaining NEs, the

collocations ‘المعالجة الكيميائية’ ‘chemotherapy’, ‘المعالجة الشعاعية’ ‘radiation therapy’, and ‘المعالجة بـ’ ‘curing by’ are related to the treatment methods entities, and the collocations ‘هذه الأعراض’ ‘these symptoms’ and ‘أعراض السرطان’ ‘the symptom of the cancer’ are related to the symptom entities.

The collocation analysis is a preparatory phase and a starting point for the next data analysis stage, which is the concordance analysis. In the concordance analysis, many of the most frequent and informative collocations are reviewed in order to extract any applicable patterns in the corpus. The collocation analysis also enhances the creation of gazetteers by providing detailed information about certain entities like the aforementioned entities in the table. The collocation analysis also produces the creation of lexical marker lists. These lexical markers are significant lexical tokens that can indicate the presences of the NEs. For example, the lexical items ‘الشعور’ and ‘ألم’ are lexical markers for the symptom entity.

Table 4.14 The most twenty informative collocations in terms of their frequency among our corpus.

Collocation	Translation	Frequency	Collocation	Translation	Frequency
ب سرطان	by cancer	242	الإصابة بـ	infection with	209
المعالجة الكيميائية	chemotherapy	108	المعالجة الشعاعية	Radiation therapy	82
معالجة سرطان	curing the cancer of	68	ابيضاض الدم	Leukaemia	57
سرطان الدرق	thyroid cancer	56	سرطان الفم	oral cancer	55
سرطان الرحم	cervical cancer	52	سرطان الكبد	Liver cancer	49
هذه الأعراض	These symptoms	48	تشخيص سرطان	Diagnosing the cancer of	46
المعالجة بـ	curing by	46	سرطان المعدة	Stomach cancer	45
سرطان الشرج	Anal cancer	44	سرطان الثدي	Breast cancer	44
ب الأشعة	By radiation	42	سرطان المرارة	gallbladder cancer	41
سرطان الكلية	Kidney cancer	40	أعراض سرطان	The symptom of the cancer	38

4.4.2.3 Concordance Analysis

Examining keyword collocates can highlight some of the most obvious themes and topics surrounding the keyword. This allows us to begin to build on the situational occurrences of the text and to see the importance of context and word order. However, collocation alone does not offer a full picture about how words function in the context of the text as a whole. Therefore, to appreciate the subtleties and different layers of meanings in a text, a concordance analysis should be used to complement the collocation analysis (Harvey, 2013). Reviewing previous research, it was soon observed that concordance analysis is a very powerful, commonly employed tool. Conducting a thorough concordance analysis is especially important when the researcher is mining a large corpus for detailed information. As our corpus extracted from the King Abdullah Health Encyclopedia is extensive as in-depth analysis of the language of the entire text, within the time limitations of the study is difficult. To overcome this problem of the corpus, a concordance analysis is carried out to investigate the meaning of words and their contextual usages.

The concordance analysis has different usages and applications. It can be used in second language education: Benzenberg (2014) found that concordance software can be used as an aid to computer-assisted language learning. He observed that “academic language classrooms are often filled with students who hail from a diverse range of disciplines who require different linguistic skills and lexical sets” (Benzenberg, 2014, p. 11). He believed that because concordance searches use authentic corpus from the first language, they may facilitate a pedagogical approach to second language learning, grounded in everyday or situational language use. Concordance analyses can also be used for extracting data from a large corpus. Liu *et al.* (2015) held that concordances are highly important to understanding the quality of given phrases in the context of a much larger document. These two uses can overlap in many ways, resulting in the development of language programmes, the creation of educational methods, the comprehension and correction of social biases, and the translation of large texts.

In our study, this allowed patterns surrounding keywords to be located and understood. Without the concordance analysis, these patterns may have gone undetected. Locating these patterns is important for extracting new keywords and NEs of interest for further analysis. Concordance analysis draws attention to other lexical items that may not be significant enough to become keywords but that otherwise form a pattern. For example, they may frequently occur with important key lexical items. They may appear more frequently alongside their synonyms, or they may pair up with other significant and infrequent lexical items to create a theme or a dialogue around certain key lexical items. These

lexical items are still very important to understanding a text but would not be noticed with a frequency-sorted lexical item list alone. The researcher is aware that it is important to be absolutely certain that enough of a sample has been extracted to support the findings. Running a concordance analysis on a smaller extract may not include all the relevant data and may not yield significant results. Concordances consist of keywords accompanied by a few words of accompanying text, which give the analyser some insight into the general context of the keyword without analysing the entire text as a whole. This makes concordance analysis particularly valuable when mining a large corpus for small, yet highly significant details. Concordance lines provide the analyser with convenient access to text samples surrounding the already isolated and categorised keywords. Unlike keyword lists and frequency-sorted lists, which focus on the quantity of lexical items used or collocations that focus largely on the quality of the lexical items relative to keywords, a concordance combines both elements, analysing both the frequency and the qualities of the keywords. This allows us to understand the meaning of keywords in the greater context of the text and even the subject (Harvey, 2013).

Concordance analysis assists in the investigation of the context of NEs of specific interest. 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 which is a freely downloadable concordance tool for Arabic corpus linguistics (Roberts *et al.*, 2006) is used to carry out the concordance analysis. Figure 4.6 illustrates the concordance analysis for the lexical item, 'سرطان' 'cancer'.

و الأمعاء الغليظة إن	سرطان	الأمعاء الدقيقة حالة نادرة
تشتمل العلامات	سرطان	الأمعاء الدقيقة على ما
المحتملة ل		
على المساعدة في	سرطان	الأمعاء الدقيقة و بيان
تشخيص		
الأكثر شيوعا ل معالجة	سرطان	الأمعاء الدقيقة و تشتمل
يقوم الأطباء ب	سرطان	الأنف عن طريق الفحوص
تشخيص		
السائل المتوي و يعتبر	سرطان	البروستات السبب الثالث
		الأكثر
غالبا ما تعتمد معالجة	سرطان	البروستات على المرحلة
		التي
و قد تشتمل أعراض	سرطان	البروستات على ما يلي
لدى الرجال المصابين ب	سرطان	البروستات لكن مستضد
		البروستات

Figure 4.6 The concordance analysis for the lexical item 'cancer', 'سرطان'.

The concordance analysis identified verb-related and noun-related patterns. Details of the extracted patterns are given in Section 4.4.3.

Sketch Engine is a corpus tool which can reveal language grammar patterns. It constructs word sketches using the usual corpus query system (CQS) functions and compares synonyms and locates 'sketch differences'. Each word sketch created by Sketch Engine is fully integrated with its relevant concordance. This means that by clicking on a collocate of interest in the word sketch, the user can be taken to a concordance of the corpus, which gives rise to that particular collocate in that grammatical relation. For example, if a user is looking at a list of high-salience objects in relation to a sketch of the verb 'spread' and clicks on 'toast', then the user will be taken to a concordance of contexts in which the noun 'toast' occurs as an object of the verb 'spread' (Kilgarriff *et al.*, 2004).

This word sketch is incredibly useful on many levels. It not only employs well-founded salience statistics and lemmatisation but also addresses other relevant questions around the keywords found in a corpus. Because a word sketch uses grammar patterns, it does not focus on an arbitrary window of text around the keyword but instead it focuses on each grammatical relation associated with this word. Currently, Sketch Engine contains 27 grammatical relations for English. After identifying a grammatical relation, the word sketch then provides a list of collocates for every grammatical relation related to a given keyword participates. For example, in the case of a verb, 'the subject, the objects, the conjoined verbs (stand and deliver, hope and pray), modifying adverbs, prepositions and prepositional objects' are all given their own list to highlight each collocate in relation to its function. Furthermore, each collocate can provide the lexicographer with a list of the corpus contexts related to the selected word and its collocate occurrences (Kilgarriff *et al.*, 2004).

As an example of the word sketch in action, Figure 4.7 shows the sketch of the word 'سرطان' 'cancer'. It shows the verb to the left, the verb to the right, the noun to the left, the noun to the right, the adjective to the left, and the adjective to the right of the word. Regular concordance lines show only the context of the word of interest. A word sketch is a more advanced tool in terms of extracting useful information out of the concordance lines. This makes a word sketch a valuable step to interpreting the concordance lines. The task of extracting useful information from concordance lines is not very accurate, and it would be easy to miss useful information. By clicking on any of these words, the Sketch Engine will show the concordance lines where this word appeared along with the word whose sketch is shown. In Figure 4.8, we can see an example of the results of selecting a related word for further analysis. Figure 4.8 is the result of selecting the first word in the column 'verb right' illustrated in Figure 4.7. This helped in studying and understanding the way the corpus is written, serving as a deeper insight into concordances and word co-occurrence. Observing and understanding these

intricacies around each NE can help the creation of a list of lexical markers. A list of lexical markers will allow for a more accurate detection of NEs and enrich the different gazetteers used.

The two highlighted lexical items in red in Figure 4.8 are the verb to the right ‘تتضمن’ ‘include’ and the lexical item ‘سرطان’ ‘cancer’, which has been sketched. Following this analysis with various keywords and NEs, Sketch Engine helps recognise entities. Sketch Engine’s lexical item analysis helps break down concordances, which further contributes to studying and understanding how the corpus is written. Through this understanding of the structure of the corpus and its writing, it will be easier to detect existing patterns associated with the NE of specific interest. By doing so, the lexical item sketch can contribute to the creation of a list of lexical markers, which are the lexical items that give an indication that the following lexical item is an entity. The end result of this would be a greater detection of entities and an enrichment of the different gazetteers that are in use.

سرطان		SaadCorpus freq = 1,021 (15,865.61 per million)									
verb_left	240 0.80	verb_right	504 1.20	noun_left	1,771 1.20	noun_right	1,353 0.90	adi_left	267 0.80	adi_right	208 0.70
يتلف	13 10.41	تتضمن	19 9.86	الرحم	99 10.35	الاصابة	135 10.95	للحاجبة	39 11.30	للتالعة	19 11.02
ان	26 10.16	كان	31 9.55	الدرق	56 9.80	معالجة	70 10.12	الدقيقة	32 10.86	المصابين	15 10.57
يسعى	5 9.26	يبدأ	18 9.49	القم	55 9.61	تشخيص	49 9.92	شائع	15 10.55	ناجحة	13 10.53
انتشر	13 9.23	بلغ	15 9.34	عنق	47 9.56	خطر	49 9.91	دائما	13 10.21	كثيرة	9 10.10
يلي	7 9.22	يعد	13 9.31	الشرح	45 9.47	اعراض	45 9.79	ثقيلي	10 10.05	مرتفعة	6 9.34
يبدأ	7 9.03	يصاب	12 9.26	القولون	45 9.46	الاعراض	43 9.55	نادرا	7 9.51	التكتيفي	5 9.05
تكون	8 8.86	يسبب	14 9.20	الكبد	48 9.44	اجل	32 8.89	الحرشفية	5 9.02	نجمية	4 9.02
يصيب	5 8.79	يصيب	12 9.20	المعدة	45 9.43	مرضى	20 8.79	رئيسي	4 8.79	سريرية	4 8.98
يمكن	15 8.65	ان	32 9.15	التدي	45 9.42	اسم	26 8.76	الحقيقي	4 8.75	الوحيدة	4 8.89
يعتمد	4 8.61	يسمى	12 9.11	المرارة	41 9.39	تيوعا	20 8.72	واحد	4 8.66	الاخرى	6 8.71
انتقل	4 8.60	يمكن	38 9.06	الجلد	44 9.37	مقدمة	19 8.72	المستخدمة	4 8.53	اخرى	6 8.10
لكن	5 8.58	ظهرت	9 9.00	الكلى	40 9.34	حالات	18 8.64	المريضة	4 8.49		
كان	9 8.38	لكن	14 8.94	المبيض	38 9.33	علاج	18 8.62				
يكون	4 7.84	يكون	16 8.90	العين	41 9.26	انواع	20 8.61				
		يؤدي	10 8.85	الغدد	38 9.26	الخاصة	17 8.58				
		يدعى	12 8.63	الفرج	37 9.26	مراحل	18 8.53				
		يتعلق	6 8.52	العظام	37 9.16	اصابة	17 8.50				
		يصابون	6 8.49	البنكرياس	34 9.16	اكتشاف	18 8.48				
		امكن	6 8.44	الرئة	32 9.11	احتمال	14 8.29				
		يعطى	6 8.41	الامعاء	35 9.06	مخاطر	13 8.22				
		تتضمن	6 8.39	القضيب	27 8.80	الوقاية	12 8.09				
		ينتشر	6 8.29	الخصية	25 8.76	خطورة	12 8.04				
		يخص	5 8.27	الكظر	26 8.74	الاتخاص	12 8.00				
		تعتمد	5 8.07	الجراحة	26 8.55	المرضى	16 8.00				
		تتضمن	4 7.88	الانف	22 8.52	ف	10 7.86				

Figure 4.7 the sketch of the word سرطان “cancer”.

4.4.3 Features Extraction step

The data analysis steps allow us to extract a set of features which can serve as cues for identifying relevant entities for our medical corpus. These features are presented to the classifier in the form of feature vector for training. In NER, features are the properties or characteristic attributes of words that a researcher feeds the computational system being used (Shalaan, 2014). A feature vector is an abstraction over the text and usually divides each word into binary values (where there are two possible results), numerical values (where there are limitless possible results) and nominal values (syntactic values) (Shalaan, 2014).

file2883791	هذا السرطان . تشتمل العلامات المحتملة ل سرطان الأمعاء الدقيقة على ما يلي : ألم بطني .
file2883791	نمو هذا السرطان . تشتمل الأعراض الشائعة ل سرطان الأمعاء الدقيقة على ما يلي : كتلة في البطن
file2883791	الصحية ل المريض . من الممكن ان تشتمل معالجة سرطان الأمعاء الدقيقة على الجراحة او المعالجة
file2883791	القولوني الورمي الغدي العائلي . تشتمل معالجة سرطان الأمعاء الدقيقة عادة على الجراحة , او المعالجة
file2883791	ان ينمو السرطان . تشتمل الأعراض الشائعة ل سرطان الأنف على ما يلي : إصابة مزمنة ل الجيوب
file2883791	العامة ل المريض . من الممكن ان تشتمل معالجة سرطان الأنف على الجراحة او المعالجة الشعاعية
file2883791	الجيوب و عدم انفتاح ها . تشتمل سيل معالجة سرطان الأنف عادة على الجراحة او المعالجة الشعاعية
file2883791	بصاها ب ه . الأعراض تشتمل الأعراض الشائعة ل سرطان الخصية على ما يلي : تغير في كيفية الشعور
file2883791	20 الى 39 عاما . تشتمل الأعراض الشائعة ل سرطان الخصية على ما يلي : تغير في كيفية الشعور
file2883791	العقد اللمفية المجاورة . تشتمل خيارات علاج سرطان الخصية غالبا على الجراحة , او المعالجة
file2883791	اية اجزاء من الجسم . يمكن ان تشتمل معالجة سرطان الرحم على الجراحة و المعالجة الشعاعية و
file2883791	الى العقد اللمفية . يمكن ان تشتمل معالجة سرطان الترح على الجراحة و المعالجة الشعاعية و
file2883791	المرحلة التي بلغ ها المرض . و قد تشتمل معالجة سرطان العين على ما يلي : الجراحة . المعالجة الاستعاعية
file2883791	الورم في خاذا الشبكية . و قد تشتمل معالجة سرطان العين على ما يلي : الجراحة . المعالجة الاستعاعية
file2883791	الصحة العامة ل المريض . قد تشتمل معالجة سرطان الكظر على الجراحة او المعالجة الشعاعية
file2883791	من اجل تشخيص سرطان الكظر . تشتمل معالجة سرطان الكظر عادة على الجراحة او المعالجة الشعاعية
file2883791	مسؤولة عن هذه الاعراض . يمكن ان تشتمل معالجة سرطان الغدد اللعابية على الجراحة و المعالجة الاستعاعية
file2883791	الانسجة القريبة مع مرور الزمن . تشتمل معالجة سرطان الفرج على ما يلي : الجراحة . المعالجة الكيميائية
file2883791	سرطانية اذا لم تجر ازالة ها . تشتمل طرق معالجة سرطان عنق الرحم عادة على الجراحة و المعالجة الشعاعية

Figure 4.8 the result of clicking on the first verb right, **تشتمل** “include”

The source of these values could be detected by the classifier through surface features, a pre-processing breakdown, the items that surround the word list, or even the characters that compose the word. It could also locate values based on a combination of the different detection methods (Oudah and Shaalan, 2013). In this section, we present the features that have been used in this study.

disease NE follows the lexical marker 'تشخيص' 'diagnosis'. Thus, using a lexical marker can enhance the NER performance.

- Stop words list

Stop word lists are a set of words within a certain text that do not carry any relevant information. These words do not bear information relevant to the application, including prepositions, pronouns, demonstratives, etc. (Elsebai, 2009). Stop word lists are widely used by the natural language processing community. For example, Samy *et al.* (2005) made use of a stop word list by using a filter to remove stop words from potential transliterated candidates. Some systems also allow the researcher to employ a blacklist, as described by Shaalan and Raza (2009), which makes it easy to discard negative evidence. A filtering mechanism can also be used to reject false matches. The use of stop words as a feature to train the classifier can help in the process of recognising the entities of special interest, although some specific entities (e.g., disease names) may not include a stop word. However, in some entities (e.g., symptoms), stop words will occur as a part of the entities. For example, the symptom 'ألم في العظام أو المفاصل' 'bone or joint pain' contains the stop words 'في' 'in' and 'أو' 'or'. A free stop word list for the Arabic language which is available online is used to check, expand, and enrich the tasks of feature extraction. A copy of the Arabic language stop words list is found in the appendix I.



Figure 4.10 The concordance lines of the lexical item diagnosis "تشخيص"

- Gazetteers

Gazetteers, otherwise known as entity dictionaries, can be highly valuable in improving NER performance (Kazama and Torisawa, 2008). They are normally made of a predefined list of NEs or keywords (Oudah and Shaalan, 2012). In most research, gazetteers are frequently utilised, and most systems developed for Arabic NER text are dependent on predefined proper name gazetteers (Alruily, 2012). Although widely used, they can be very difficult and sometimes time-consuming to build. This is because many Arabic resources, such as corpus and other gazetteers, can be difficult to access and expensive to purchase. Those that are not out of reach or budget are few and may still incur a cost (Alruily, 2012). The use of gazetteers could improve the overall precision of the NER system. However, reliance only on gazetteers would affect the overall recall of the NER system. For our system, four different gazetteers were manually built in which each gazetteer is related to a specific named entity.

- The entity type tag

As explained in Section 4.4.1.3, four NE classes were recognised and labelled with the following tags: disease name (D), symptoms (S), treatment methods (T), and diagnosis methods (G). The types of entity tags of the target word and the two preceding and two following words (\pm two-word sliding window) are used as features to train the classifier.

- The definiteness (existences of ‘AL’)

‘AL’ is the definite article of the Arabic language, equivalent to the English definite article ‘the’. The presence of ‘Al’ renders a noun definite. Unlike the English equivalent, it occurs as a prefix particle that is affixed to the noun. For example, كتاب kitāb, meaning ‘book’, can be made definite by adding the prefix ‘AL’, which transforms it into الكتاب al-kitāb, meaning ‘the book’. The definiteness feature has five values that indicate the presence of the definite article ‘AL’. These are indefinite (i), definite (d), construct/poss/idafa (c), not applicable (na), and undefined (u). The definiteness of the target word and the two preceding and two following words (\pm two-word sliding window) were used as features to train the classifier. From observing the data, it seems that there is a hidden structure in the way the Arabic text is written. This is because of the nature of genitive nouns in the Arabic language. Usually, the genitive nouns are a compound of two words in which the first word is indefinite (with no ‘AL’ article), and the second word is definite (with the ‘AL’ article). By surveying our corpus, most of the disease NEs are represented as genitive nouns in which the first indefinite word is ‘سرطان’ ‘cancer’. Figure 4.11 shows some examples of these disease NE genitive nouns.

file2883791	الكميائية او غير ها من العوامل . و في حالة	سرطان	الجلد , تكون اشعة الشمس هي المسؤولة عن
file2883791	يختفي خلف الاعضاء الاخرى . و قد تكون معالجة	سرطان	البنكرياس صعبة ب سبب التأخر في اكتشافه
file2883791	الخلاصة ابيضاض الدم هو نوع معروف من انواع	سرطان	الدم . و هناك انواع مختلفة من ابيضاض الدم
file2883791	المحيطة ب ها قدرة على المساعدة في تشخيص	سرطان	الامعاء الدقيقة و بيان ما اذا كان قد انتشر
file2883791	التجريف في طريق ه الى الحلق عند التنفس . ان	سرطان	التجريف الانفي و الجيوب الانفية حالة نادرة
file2883791	التعاضدية و المعالجة الكيماوية . مقدمة يبدأ	سرطان	الاحف في خلايا التجريف الانفي او الجيوب
file2883791	هذا البرنامج سرطان الامعاء الدقيقة , و ليس	سرطان	الثولون . يبدأ سرطان الامعاء الدقيقة في

Figure 4.11 Some examples of disease NE genitive nouns.

- Patterns (local grammar)

This feature is not limited to the window size but is related to the whole sentence. The concordance and word sketch analyses assist in the task of identifying the hidden patterns in our corpus. Two different types of patterns are identified: verb-related patterns and noun-related patterns, described below. A total of 35 patterns are extracted from our corpus.

- Verb-related patterns (Figure 4.12)

a. The verb “include” تتضمن

Pattern 1: تتضمن “include” + الأعراض “the symptoms” + : colon + Symptom NEs

Example 1	تتضمن الأعراض ل سرطان الرئة : السعال الذي لا يشفى و يتفاقم مع مرور الوقت
Translation	Symptoms of lung cancer include: - A cough that doesn't go away
Example 2	تتضمن الاعراض ما يلي : الشعور ب الثقل في الحوض . الالم اسفل البطن
Translation	The symptoms include: heavy feeling in pelvis. Pain in lower abdomen

Pattern 2: تتضمن “include” + الأعراض “the symptoms” + Symptom NEs

Example 1	تتضمن الأعراض الما او تورما او كتلا في الخصيتين او المنطقة الاربية
Translation	Symptoms include pain, swelling or lumps in your testicles or groin area
Example 2	تتضمن هذه الاعراض اصفرار الجلد و العينين , و الما في البطن و الظهر
Translation	The symptoms include yellowing of the skin and eyes, pain in the abdomen and back.

b. The verb “include” يتضمن

Pattern 3: “include” يتضمن + “the treatment” العلاج : colon + treatment method NEs

Example	قد يتضمن العلاج : العلاج الكيميائي و العلاج ب الاشعة و العلاج ب العمل الجراحي
Translation	Treatment may include chemotherapy, radiation and surgery

c. The verb “occur” يحدث

Pattern 4: “occur” يحدث + “cancer” سرطان + disease NE

Example 1	يحدث سرطان عنق الرحم ب سبب انواع متعددة من الفيروسات
Translation	Cervical cancer is caused by several types of a virus
Example 2	يحدث سرطان المبيض عند النساء اللواتي تجاوزن الخمسين من العمر غالبا
Translation	Ovarian cancer usually happens in women over age 50

d. The verb “spread” ينتشر

Pattern 5: “spread” ينتشر + “cancer” سرطان + disease NE

Example 1	من الشائع أن ينتشر سرطان الكظر الى العظام و الكبد و الرئتين و غشاء البريتوان
Translation	Adrenal gland cancer commonly spreads to the bones, liver, lungs, and peritoneum.
Example 2	يمكن أن ينتشر سرطان الكبد الى الرئتين و العظام و العقد اللمفية المجاورة ل الكبد
Translation	Liver cancer can spread to the lungs, bones, and lymph nodes near the liver

e. The verb “infect” يصيب

Pattern 6: “infect” يصيب + “cancer” سرطان + disease NE

Example 1	قد يصيب سرطان الثدي اكثر من سيدة في العائلة الواحدة
Translation	Breast cancer may involve more than one member of a family
Example 2	يصيب سرطان الفم اي جزء من جوف الفم , ب ما في ذلك الفم و الشفتان
Translation	Oral cancer is cancer that develops in any part of the oral cavity, including the mouth and lips.

f. The verb يشعر “feel”

Pattern 7: يشعر “infect” + the preposition بـ “ba” + symptoms NEs

Example 1	و يشعر بـ الوهن و التعب
Translation	and feel weak and tired
Example 2	يشعر المريض عندئذ بـ ألم شديد في البطن و غثيان
Translation	The patient will have severe pain in the abdomen area, with nausea.

g. The verb ينشأ “arise”

Pattern 8: ينشأ “arise/start” + سرطان “cancer” + disease NE

Example 1	يمكن ان ينشأ سرطان العظم ثم ينمو عبر الطبقة الخارجية لـ العظم
Translation	a bone tumor may grow through the bone’s outer layer
Example 2	يمكن ان ينشأ سرطان الرحم ثم ينمو عبر الطبقة الخارجية لـ الرحم
Translation	Cervical cancer begins on cells on the surface of the cervix

ii. Noun-related patterns

The discovered patterns were divided to four different types:

a. The disease NE patterns (Figure 4.13)

○ The noun اكتشاف “discovery/detection”

Pattern 1: اكتشاف “discovery/detection” + سرطان “cancer” + disease NE

Example 1	إن من الصعب اكتشاف سرطان البنكرياس مبكراً لأنه لا يسبب اعراضاً
Translation	Pancreatic cancer is hard to catch early. It doesn't cause symptoms right away.
Example 2	يزيد اكتشاف سرطان الغدد اللعابية في وقت مبكر من فرص نجاح معالجته
Translation	Detecting salivary gland cancer early increases the chances of a successful treatment.

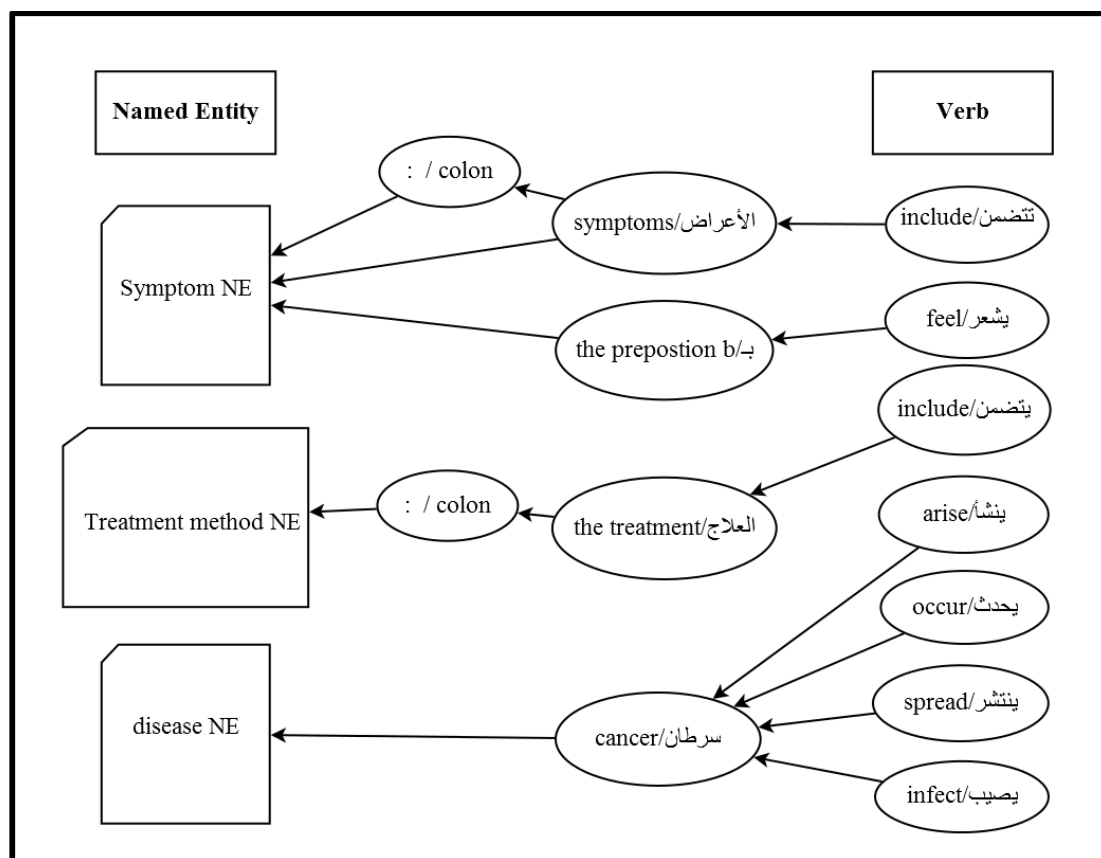


Figure 4.12 The verb-related patterns.

○ The noun انتشار “spread/ prevalence”

Pattern 2: انتشار “spread/ prevalence” + سرطان “cancer” + disease NE

Example 1	يظهر التصوير المقطع ب الاصدار البوزيتروني انتشار سرطان الكبد الى اي جزء من الجسم
Translation	A CT scan can show if the liver cancer has spread to the lungs
Example 2	ان التصوير الشعاعي ل الصدر الذي يظهر اعضاء الجسم و عظامه مفيد ايضا في اكتشاف انتشار سرطان المرارة
Translation	A chest x-ray of the organs and bones inside the chest is helpful in showing if gallbladder cancer has spread.

○ The noun تشخيص “diagnosis”

Pattern 3: تشخيص “diagnosis” + سرطان “cancer” + disease NE

Example 1	يمكن إجراء خزعة من أجل تشخيص سرطان العظام
Translation	A biopsy may be performed to diagnose bone cancer.
Example 2	تصوير الاوعية الفلوروسيني قادر ايضا على تشخيص سرطان العين
Translation	A fluorescein angiography can also diagnose eye cancer.

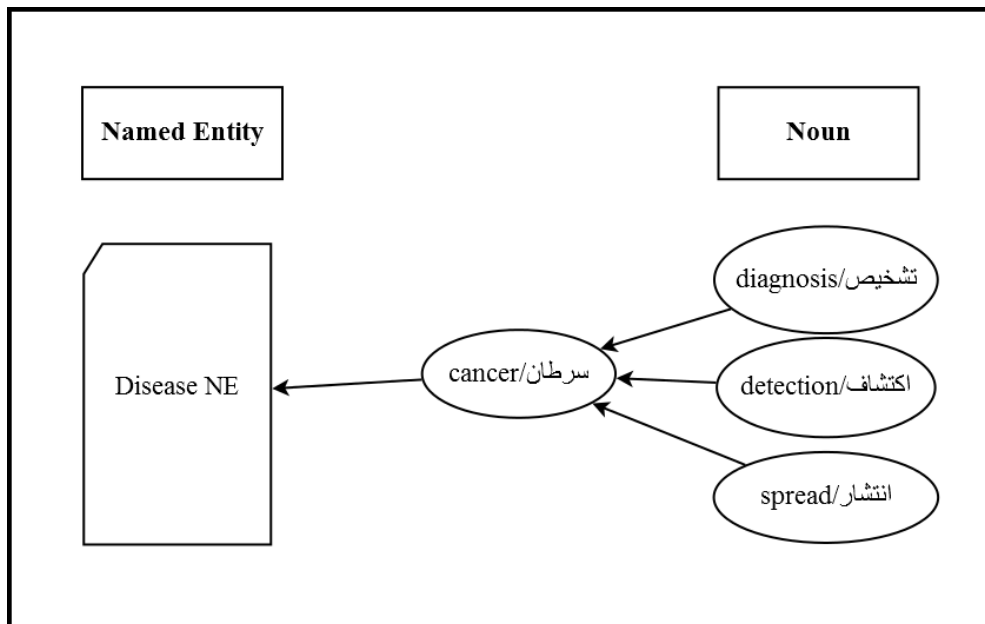


Figure 4.13 The disease NE noun-related patterns.

b. The symptom NE patterns (Figure 4.14)

○ The noun الشعور “the feeling”

Pattern 1: الشعور “the feeling” + ب “the preposition b” + symptom NEs

Example 1	من العلامات الاخرى : الوهن و الشعور ب التعب . الغثيان و القيء
Translation	Other signs include: Weakness or feeling very tired. Nausea or vomiting.
Example 2	نقص الشهية والشعور ب الإمتلاء
Translation	Loss of appetite and feelings of fullness.

Pattern 2: الشعور “the feeling” + ب “the preposition b” + symptom NEs + و/أو (and/or) + symptom NEs

Example 1	الشعور بـ ب التخمة او النفخة بعد وجبة صغيرة
Translation	Feeling full or bloated after a small meal.
Example 2	الشعور بـ الضعف و الخدر في الوجه
Translation	Feeling of numbness or weakness in the face

- The noun شعور “feeling”

Pattern 3: “the preposition ب” + “feeling” + symptom NEs

Example 1	يمكن ان يتسبب سرطان الخصية ايضا بـ ما يلي : شعور بـ ثقل في الصفن
Translation	Testicular cancer can also cause: A feeling of heaviness in the scrotum.
Example 2	وجود كتلة او شعور بـ الثقل في القسم العلوي من البطن
Translation	A lump or feeling of heaviness in the upper abdomen.

- The noun ألم “pain”

Pattern 4: ألم “pain” + في “the preposition in” + symptom NE

Example 1	ان الاعراض المألوفة ل سرطان المعدة هي : انزعاج او ألم في منطقة المعدة
Translation	Common symptoms of stomach cancer are: Discomfort or pain in the stomach area.
Example 2	يمكن ان تظهر لدى مريض سرطان الفم الاعراض التالية ايضا : صعوبة او ألم في البلع
Translation	Common symptoms of oral cancer may also include: Difficulty or pain when swallowing.

Pattern 5: ألم “pain” + عند “the adverb when” + symptom NE

Example 1	أعراض سرطان الفم هي : ظهور بقع بيضاء او حمراء في الفم . مشاكل او ألم عند البلع
Translation	Symptoms of oral cancer include: - White or red patches in your mouth. Problems or pain with swallowing.
Example 2	ان الاعراض المألوفة ل السرطان المريئي هي : انحشار الطعام في المريء , و قد يعود ادراجه الى الخلف . . ألم عند البلع
Translation	Common symptoms of esophageal cancer are: Food getting stuck in the esophagus, and food may come back up. Pain when swallowing.

Pattern 6: ألم “pain” + خلال “the adverb during/through” + symptom NE

Example 1	هناك اعراض شائعة اخرى ل سرطان عنق الرحم , و منها : الم حوضي . الم خلال الجماع
Translation	Other common symptoms of cervical cancer include: Pelvic pain. Pain during sex.

Pattern 7: ألم “pain” + فوق “the adverb *above*” + symptom NE

Example 1	و من اعراضه : اليرقان (اصفرار الجلد و بياض العينين) . الم فوق المعدة
Translation	Symptoms include: - Jaundice (yellowing of the skin and whites of the eyes) - Pain above the stomach.

○ **The noun** وجود “existence”

Pattern 8: وجود “existence” + كتلة “lump” + في the preposition in+ symptom NE

Example 1	لكن العرض الأكثر شيوعا ل سرطان الغدد اللعابية هو وجود كتلة في منطقة الاذن او الوجنة او الفك او الشفة او في داخل الفم
Translation	The most common symptom of salivary gland cancer is a lump in the area of the ear, cheek, jaw, lip, or inside the mouth.
Example 2	من الاعراض الشائعة ل سرطان الكبد : وجود كتلة او شعور ب الثقل في القسم العلوي من البطن
Translation	Some common symptoms of liver cancer are: A lump or feeling of heaviness in the upper abdomen.

Pattern 9: وجود “existence” + كتلة “lump” + قرب near + symptom NE

Example 1	الاعراض الاعراض الشائعة ل سرطان الشرج هي :وجود كتلة قرب الشرج
Translation	Common symptoms of anal cancer are: A lump near the anus

○ **The noun** ظهور “appearance”

Pattern 10: ظهور “appearance” + كتلة “lump” + في the preposition in+ symptom NE

Example 1	من الممكن ان يسبب سرطان الانف ايضا ظهور كتلة او تقرح لا يشفى في داخل الانف
Translation	Nasal cancer may also cause a lump or sore inside of the nose that does not heal.
Example 2	الاعراض ان الاعراض الشائعة ل سرطان الكظر تشتمل على ما يلي : ظهور كتلة في البطن
Translation	Common symptoms of adrenal gland cancer include: A lump in the abdomen.

Pattern 11: ظهور “appearance” + كتلة “lump” + على the preposition on+ symptom NE

Example 1	ظهور كتلة دهنية على الجهة الخلفية من الرقبة
Translation	A lump of fat on the back of the neck

Pattern 12: ظهور “appearance” + كتلة “lump” + ب the preposition b+ symptom NE

Example 1	ان العرض الأكثر شيوعا ل سرطان الغدد اللعابية هو ظهور كتلة في المنطقة المحيطة ب الاذن
Translation	The most common symptom of salivary gland cancer is a lump in the area of the ear.

○ **The noun** احمرار “redness”

Pattern 13: احمرار “redness” + في the preposition in+ symptom NE

Example 1	يؤدي الورم الارومي الشبكي الى ظهور اعراض مختلفة , من بينها : الم او احمرار في العين
Translation	Retinoblastoma causes different symptoms, including: Pain or redness in the eye.

○ **The noun** حرقة “heartburn”

Pattern 14: حرقة “Heartburn” + في the preposition in+ symptom NE

Example 1	يمكن ان يشكو مريض سرطان المعدة ايضا من : حرقة في راس المعدة او عسر الهضم
Translation	Symptoms of stomach cancer also include: Heartburn in top of stomach or indigestion

○ **The noun** قرحة “ulcer”

Pattern 15: قرحة “ulcer” + في the preposition in+ symptom NE

Example 1	اعراض سرطان الفم هي : ظهور بقع بيضاء او حمراء في الفم . قرحة معنزة في الفم
Translation	Symptoms of oral cancer include: White or red patches in your mouth. A mouth sore that won't heal.

c. The treatment methods NE patterns (Figure 4.15)

○ **The noun** استخدام “the usage”

Pattern 1: استخدام “the usage” + العلاج “the treatment” + ب “the preposition b” + treatment method NEs

Example 1	يجري استخدام العلاج ب الاشعة مع المعالجة الكيميائية في بعض انواع ابيضاض الدم
Translation	Radiation therapy is used with chemotherapy for some kinds of leukemia.

Example 2	يمكن أيضا استخدام العلاج ب الاشعة ل معالجة سرطان البنكرياس
Translation	Radiation therapy can also be used to treat pancreatic cancer.

Pattern 2: استخدام “the usage” + المعالجة “the treatment” + treatment method NEs

Example 1	يجري استخدام العلاج ب الاشعة مع المعالجة الكيميائية في بعض انواع ابيضاض الدم
Translation	Radiation therapy is used with chemotherapy for some kinds of leukemia.
Example 2	يمكن استخدام المعالجة الهرمونية لدى النساء في المراحل المتقدمة من سرطان الرحم
Translation	Hormone therapy may be used for women with advanced uterine cancer.

Pattern 3: استخدام “the usage” + جرعات “doses” + من the preposition “of” + treatment method NEs

Example 1	يجري استخدام جرعات منخفضة جدا من الاشعة السينية
Translation	Today, x-rays use very low doses of radiation
Example 2	حيث يتم ازالة نقي العظم الذي يسبب ابيضاض الدم عند المريض ب استخدام جرعات عالية من الادوية و الاشعة
Translation	the patient's leukemia-producing bone marrow is destroyed by high doses of drugs and radiation

○ The noun المعالجة “the treatment”

Pattern 4: المعالجة “the treatment” + ب “the preposition b” + treatment method NEs

Example 1	ان المعالجة ب هرمون الدرق يمكن ان تبطئ نمو خلايا سرطان الدرق المتبقية في الجسم بعد الجراحة
Translation	Thyroid hormone treatment can slow the growth of thyroid cancer cells left in the body after surgery.
Example 2	المعالجة ب التبريد قادرة ايضا على التعامل مع سرطان العين
Translation	Cryotherapy can also treat eye cancer.

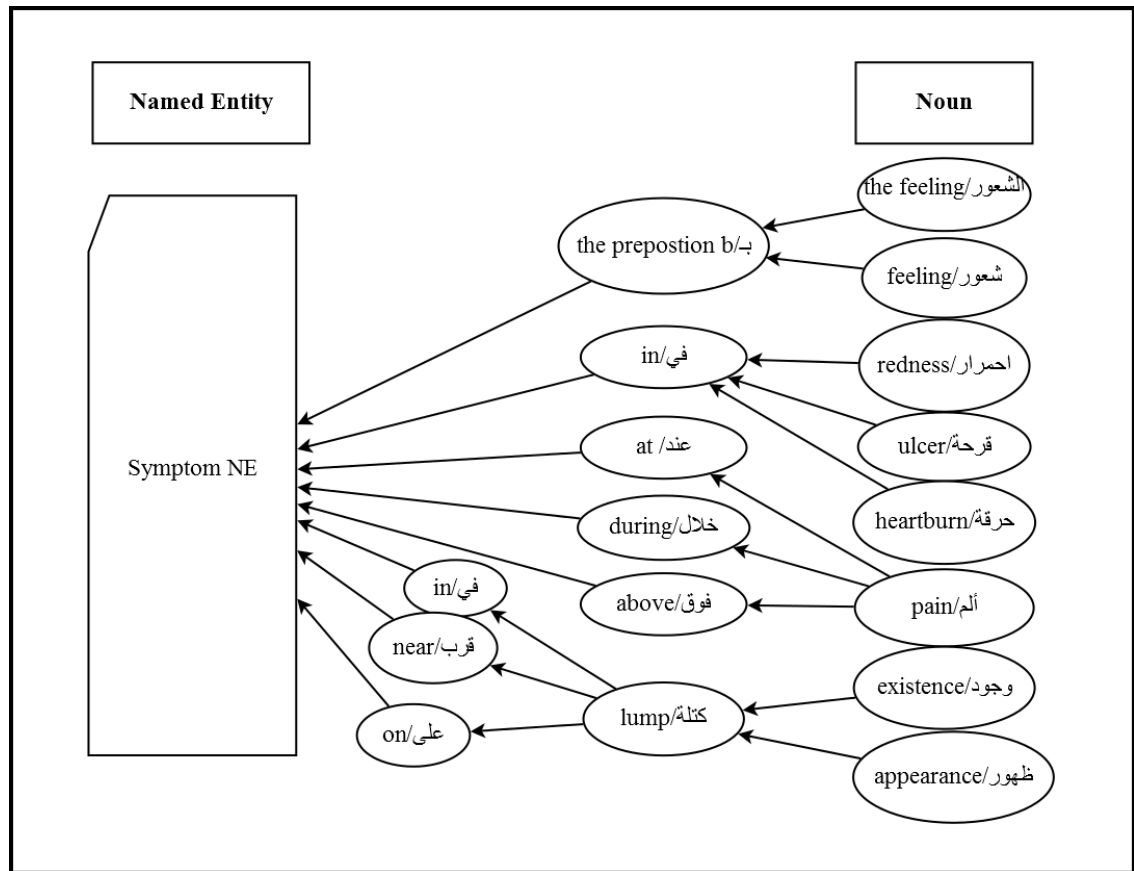


Figure 4.14 The symptom NE noun-related patterns.

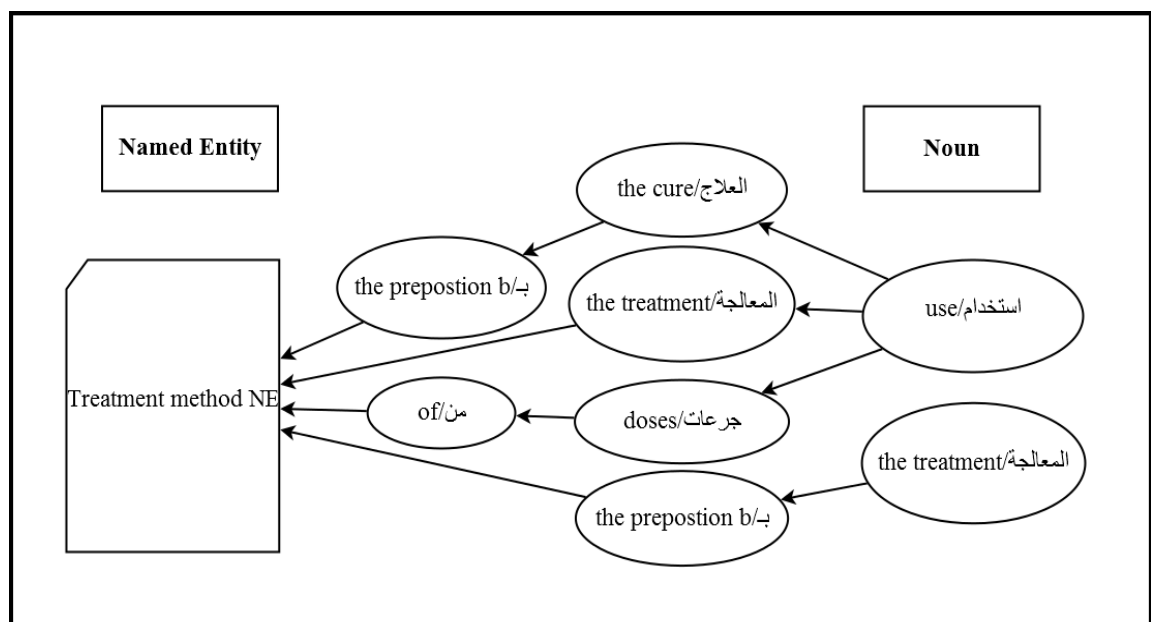


Figure 4.15 The treatment methods NE noun-related patterns.

d. The diagnosis methods NE patterns (Figure 4.16)

- The noun استخدام “the usage”

Pattern 1: استخدام “the usage” + فحوص “tests” + diagnosis method NEs

Example 1	من الممكن استخدام فحوص الدم من أجل قياس مدى سلامة عمل الكبد
Translation	Blood tests can be used to check for liver problems.
Example 2	يجري استخدام فحوص تصويرية غالبا من أجل تحديد المرحلة التي بلغها السرطان
Translation	Imaging tests are often used to determine the stage of the cancer.

- The noun التصوير “the scanning/the imaging”

Pattern 2: التصوير “the scan/imaging” + ب “the preposition b” + diagnosis method NEs

Example 1	يمكن أن تكتشف الأورام الخفية ب واسطة التصوير ب الرنين المغناطيسي أيضا
Translation	A CT scan may need to be used to check for hidden tumors.
Example 2	يستطيع التصوير ب الإصدار البوزيتروني أن يظهر ما إذا كان سرطان المعدة قد انتشر إلى أماكن أخرى من الجسم
Translation	A PET scan can show if the stomach cancer has spread elsewhere in the body.

- The noun تصوير “scanning/imaging”

Pattern 3: تصوير “scan/imaging” + ب “the preposition b” + diagnosis method NEs

Example 1	يمكن إجراء تصوير ب الأمواج فوق الصوتية من أجل تشخيص سرطان الرحم
Translation	An ultrasound may also be done to diagnose uterine cancer.
Example 2	يمكن إجراء تصوير ب الأشعة السينية ل تحديد مرحلة سرطان الفم
Translation	An x-ray is one test that may be done to determine the stage of oral cancer.

- The noun الفحص “the test/examination”

Pattern 4: الفحص “the test/examination” + ب “the preposition b” + diagnosis method NEs

Example 1	تتضمن الفحوص التي تساعد على تشخيص سرطان الخصية : الفحص ب الامواج فوق الصوتية
Translation	Tests that can help diagnose testicular cancer include: Ultrasound exam.
Example 2	هناك فحوصات او اختبارات اخرى ل تشخيص الاصابة ب سرطان القولون : حيث يوجد الفحص ب التصوير المقطعي المحوسب
Translation	Other tests may also be done to diagnose colon cancer. One such test is called a CAT scan.

○ The noun صور “images”

Pattern 5: صور “images” + ب “the preposition by” + diagnosis method NE

Example 1	حيث تؤخذ صور ب الاشعة السينية ل مريء و معدة المريض بعد ان يشرب محلول الباريوم
Translation	After you drink a barium solution, you have x-rays taken of your esophagus and stomach.
Example 2	يجري النقاط صور ب الاشعة السينية بحثا عن اية تغيرات في الوريد
Translation	Then x-rays are taken to see if there are any changes to the vein.

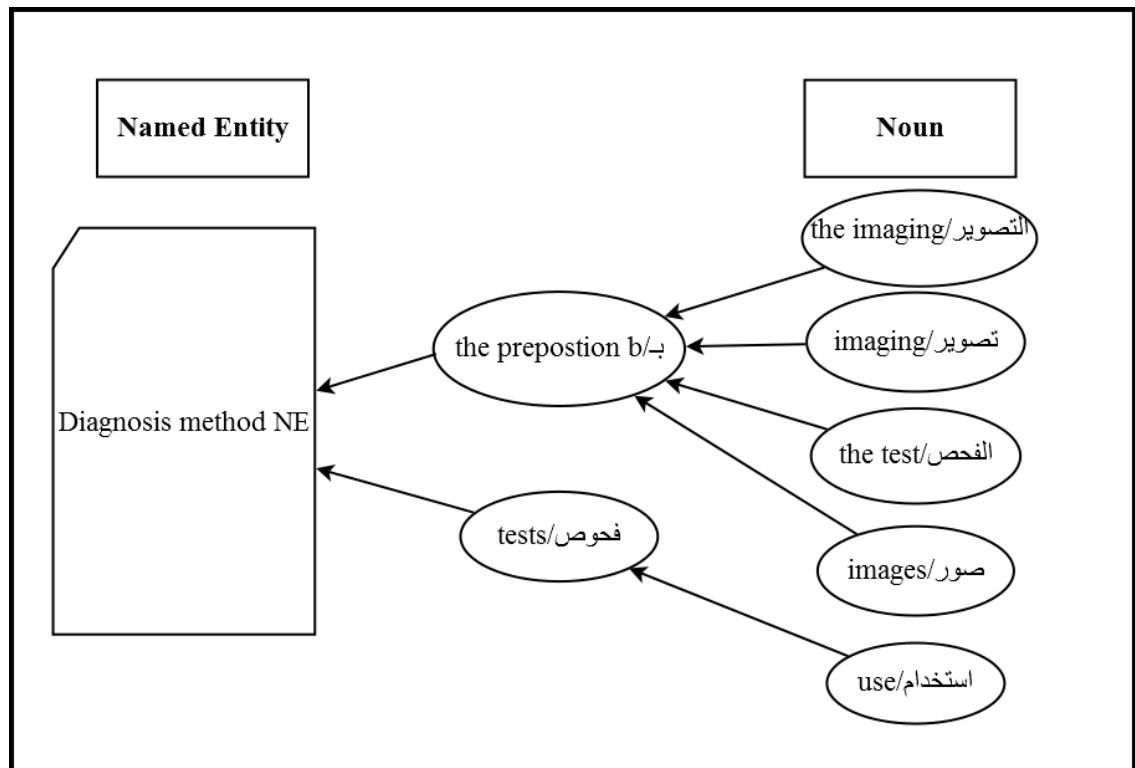


Figure 4.16 The diagnosis method NE noun-related patterns.

4.5 Summary

This chapter has described the language processing steps necessary to analyse our Arabic medical corpus prior to the extraction of named entities. A summary of the important findings and issues are listed below.

Our proposed approach to named entity recognition system is applied to the medical corpus which was extracted from King Abdullah Bin Abdulaziz Arabic Health Encyclopedia (KAAHE); the selected domain for the analysis is related to the domain of cancer which is of great concern in Saudi Arabia. The first stage of NAMERAMA involved a series of NLP tasks and produced a set of features and pattern to be employed by BBN in the second stage. Two important results are achieved. First the corpus was tokenised using AMIRA tool, POS tagged using MADAMIRA tool, and annotated manually. Second, AMIRA tool tokeniser has performed well with 91.3%, 88.5%, and 89.9% for precision, recall, and F-measure, respectively. However, number of challenges has been encountered during the tokenisation step, specifically when dealing with lexical items starting with the letter *A* after ‘*AL*’ determiner, conjunction ‘*b*’ and ‘*w*’ due to the morphological structure of the Arabic language. In spite of these difficulties, the AMIRA POS tagger has achieved an accuracy of 84% when it was applied on small set of data (5,119 tokens). The AMIRA POS tagger performs less favourably than English parsers in the area of adverbs, adjectives and genitive nouns, and, in particular, broken plurals with 32% errors. The annotation of our medical corpus is another challenging task and has required a set of annotation guidelines to be considered in order to achieve consistency. These annotation guidelines were described in Section 4.4.1.3.

Finally, the three data analysis techniques, namely frequency analysis, collocation analysis, and concordance analysis, have been used to extract relevant features for the recognition of our domain specific entities. The data analysis task has helped us study the language used in the medical domain and identify key characteristics and meaningful features namely, gazetteers, lexical markers, patterns, stopwords, and definiteness. However, because relying solely on data analysis methods to extract and select features can mislead the classifier, further efforts have been required to evaluate and rank these features before feeding these into BBN; this is discussed further in the next chapter.

Chapter 5: Bayesian Named Entity Network

5.1 Introduction

This chapter describes the second stage of NAMERAMA which relates to the classification and recognition of appropriate named entities. There are three major steps: data transformation, feature ranking, and Bayesian network implementation. The first part of this chapter describes the process of converting the corpus into a feature vector to be readable by the Bayes Server tool. The second part of this chapter explains the feature ranking approach using the likelihood ratio and Naive Bayes network. The third part discusses the implementation of Bayesian Belief Network (BBN) approach to the classification and recognition of the appropriate named entities. The impact of the sliding window on the classification and recognition performance is evaluated and a five-fold cross-validation experiment on the corpus for each class of named entities is discussed in the last section.

5.2 Data Transformation Step

The goal of the data transformation step is to transform the data to be readable by the Bayesian belief network tool. Prior to this step, the outputs from the natural language processing stage are stored in different files, including:

- *The textual dataset*: consist of 27 text files and 62,504 tokens.
- *The annotated 27 text files*: comprise the entity type of each token in the data.
- *The POS tagged data*: 27 text files store the POS tag of each token in the data.
- *Stopwords file*: lists the stopwords.
- *Lexical marker files*: contain the lexical markers related to a specific named entity category.
- *Gazetteers*: consist of four different text files that comprise the common NEs among the data. Each gazetteers file is related to a specific named entity category.
- *27 Definiteness files*: comprise the definiteness (the existences of ‘*Al*’ article in a lexical item) feature for our data.
- *Patterns file*: comprises different patterns that are related to the specific named entities.

In the data transformation step, relevant data are selected, transformed and consolidated into forms appropriate for analysis and training by the Bayesian network classifier. The output of the data transformation step is a single feature vector file that contains each token in our data along with its features.

5.3 Feature Ranking

As Shalaan (2014) explains, features in NER are important properties or characteristic attributes of words which are used to train a given classifier being used. Generally, there are two primary approaches to tackle NER task, namely rule-based approach and machine learning approach. Rule-based systems have to rely on hand crafted rules extracted from specific domain experts, and even then high recall can be very difficult to achieve. Furthermore, they are also not portable, are expensive and hard to maintain (Baluia *et al.*, 2000) and tend to focus on extracting entities such as organisations, persons and locations, or temporal expressions such as dates and times, or numerical quantities such as currency and percentages which can be identified more easily than complex entities such as disease names, symptoms, diagnosis and treatments, which have no explicit recognisable expressions or markers. Machine Learning (ML) systems can achieve good performance by learning from features or patterns but demand large computational resources. Their success depends on identifying a set of features from which a classification model is constructed that can adequately represent the corpus being used (Hall, 1999). This study applies machine learning, which is a supervised learning algorithm. The training set, which is based on the annotated corpus from the natural language stage, captures the entities and their associated feature types (e.g. gazetteers, lexical marker, stopwords, POS tags, patterns and definiteness). A window of up to five words, including two words preceding and two words following the current word is adopted.

Two methods are used to rank and evaluate the importance of each feature with respect to the correct named entities: the likelihood-ratio and a simple Naïve Bayesian Network (NBN), using Bayes server which is a tool for modelling Bayesian networks.

5.3.1 Likelihood-ratio

The likelihood-ratio is a statistical significance test aimed at deciding which features are likely to influence a target variable of interest. This is applied on the target variables in our data, which consist of five different values (disease, symptom, treatment method, diagnosis method, and not an entity). When evidence is entered in a Bayesian network the probability or likelihood of that evidence (e), denoted $p(e)$, is computed by the tool. This probability, $p(e)$, indicates how likely it is that the network has generated that data. The likelihood ratio for a given evidence (e) is given by the ratio of the probabilities of the evidence $p(e)$ occurring given that the statement is either true or false. The result of this test shows that all the features are checked as true features. This may be caused by the fact that 11% of the values of the target variable in our corpus are labelled *as named entities* (i.e. disease,

symptom, treatment method, diagnosis method) *consequently* the likelihood-ratio is overwhelmed by the NotEntity values in the data.

5.3.2 Naïve Bayes network

Bayesian classifiers are statistical classifiers. They can predict the probability of class membership, for example, the probability that any sample belongs to any given class. These classifiers are based on Bayes' theorem. A naïve Bayesian classifier assumes that the effect of an attribute value on a given class operates independently of the values of other attributes. This assumption is referred to as class conditional independence. It allows the researcher to simplify the computation of the classifier and that is why it is named *naïve* (Leung, 2007). While naïve Bayes may be a simple tool, it is invaluable for data classification. It is employed in a wide range of classification tasks; it is employed for highly varied data, including medical texts, computer network data, and text recognition. It is simple, yet produces solid and reliable results from its classifications. It is efficient and easy to construct due to the assumption that all the features are independent of each other (Amor *et al.*, 2004) as shown in Figure 5.1. Consequently, the cost of joint probability factorisation is reduced as much as possible to its simplest form. However, when applied to real datasets, the independence assumption can prove impractical and result in inaccuracies.

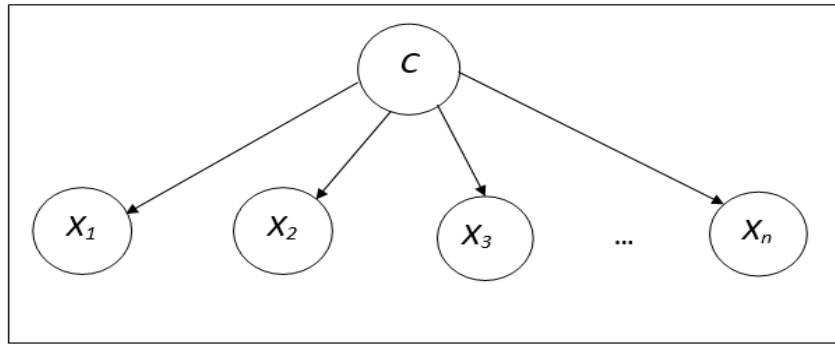


Figure 5.1 A typical Naïve Bayes network (Ang *et al.*, 2016).

An example of our simple Naïve Bayes network structure is given in Figure 5.2 where each feature is expressed in terms of window size of a five-word, and each feature node includes a set of states representing the possible values of this feature. For each feature, our simple Naïve Bayes network is implemented four times once for each entity type described below.

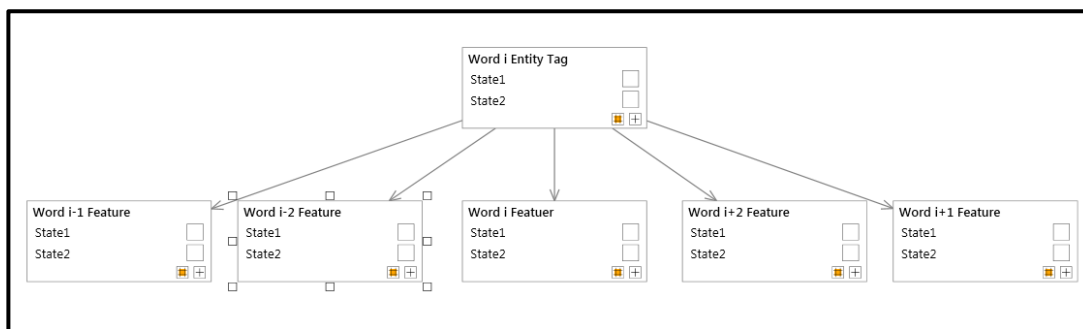


Figure 5.2 A simple Naïve Bayes network to evaluate the features.

- Named Entity: Disease

The simple Naïve Bayes network structure is applied to evaluate and rank the features performance in order to recognise the named entity, disease. Table 5.1 presents the results which show that gazetteer features of the previous two words, current word, and following two words yield a true positive count of 436, representing 89.53% precision, 81.95% recall, and an F-measure of 85.57%. The misclassification errors for this entity include 51 false positives, where non-entities are flagged as entities, and 96 false negatives, where entities are overlooked. The lexical marker features for the same five-word window has a true positive count of 392, with precision of 98.99%, recall of 73.68%, and an F-measure of 84.48%.and include 4 false positives and 140 false negatives. The annotation tags have produced 149 true positives versus 474 false positives and 383 false negatives, accounting for the worst precision rate for this disease entity type at only 23.92%. Recall is at 28.01%, and the F-measure is at 25.80%. Definiteness features show little influence, with 97 true positives, 192 false positives, and 435 false negatives, yielding 33.56% precision, 18.23% recall, and an F-measure of 23.63%. The two poorest categories in terms of recall are the pattern features and POS tags, although both show high precision. Pattern features yield 30 true positives, no false positives, 502 false negatives, giving them the highest possible precision rate at 100%, a poor recall rate of 5.64% and a poor F-measure of 10.68%. The POS tags generate only 17 true positives, 31 false positives, and 515 false negatives, which is the highest false negative count for this disease entity table. This caused a 35.42% precision rate, but the lowest recall at 3.20%, also yielding the poorest F-measure at 5.86%. The reason for the high F-measure achieved by the gazetteer features is due to the fact that most disease entities in the corpus contain the word *cancer*, which is listed in the gazetteers and lexical markers indicating that the following words are disease entities. The results of applying the simple NBN using only the annotation tags show us that entities tend to appear next to each other and hence, knowing whether or not the previous or following two words are disease entities may help our network to detect whether the current word is or is not an entity.

Table 5.1 The results of applying the NBN using each feature for the category of disease entities.

No	Features	TP	FP	FN	Precision	Recall	F-measure
1	Gazetteers features	436	51	96	89.53%	81.95%	85.57%
2	Lexical marker features	392	4	140	98.99%	73.68%	84.48%
3	The annotation tags	149	474	383	23.92%	28.01%	25.80%
4	Definiteness features	97	192	435	33.56%	18.23%	23.63%
5	The pattern features	30	0	502	100 %	5.64%	10.68%
6	POS tags	17	31	515	35.42%	3.20%	5.86%

The ability of our simple NBN using POS tag features to recognise entities is somewhat limited on their own. However, using POS tag features alongside other features may improve the performance of the network. The simple NBN using only the definiteness features achieved a 23.64% F-measure. This might indicate that there is a hidden pattern in how the *al* article is used among the disease entities.

- Named Entity: Diagnosis Methods

Table 5.2 presents the results related to the entity diagnosis methods. As with disease, the gazetteer features represent the most accurate feature, with 102 true positives, 23 false positives, and 151 false negatives. This gives a precision rate of 81.60%, a recall rate of 40.32% and an F-measure of 53.97%. Annotation tags yield a total 149 true positives, 168 false positives, and 104 false negatives, providing 47.00% precision, 58.98% recall, and 52.28% F-measure. Lexical markers return 58 true positives, 3 false positives, and 195 false negatives, at 95.08% precision, 22.92% recall, and 36.94% F-measure. Pattern features show a low score of 1 true positive and 2 false positives, with 252 false negatives. This gives a precision rate of 33.33%, a recall rate of 0.40%, and a total F-measure of 0.78%. The worst score was again related to the definiteness feature, providing no positives and unreadable results.

Table 5.2 The results of applying the NBN using each feature for the diagnosis method entity category.

No	Features	TP	FP	FN	Precision	Recall	F-measure
1	Gazetteers features	102	23	151	81.60%	40.32%	53.97%
2	The annotation tags	149	168	104	47.00%	58.89%	52.28%
3	Lexical marker	58	3	195	95.08%	22.92%	36.94%
4	The pattern features	33	13	220	71.74%	13.04%	22.07%
5	POS tags	1	2	252	33.33%	0.40%	0.78%
6	Definiteness features	0	0	253	N/A	0.00%	N/A

- Named Entity: Treatment Methods

Table 5.3 presents the results for the treatment methods entity which generate some of the poor scores out of all the NBN structures. The highest influential feature is the gazetteers, with 185 true positives, 141 false positives, and 108 false negatives returning 56.75% precision, 63.14% recall, and a total F-measure of 59.77%. The second influential feature is the annotation tags, with 121 true positives, 205 false positives, and 172 false negatives, yielding a precision rate of 37.12%, a recall of 41.30%, and a total F-measure of 39.10%. From there on, we see a sharp decline in the influence of the remaining features, these include i) lexical marker features with 14 true positives, 3 false positives, and 279 false negatives, resulting in 82.35% precision, 4.78% recall and 9.03% F-measure, ii) pattern features with 10 true positives, 2 false positives, and 283 false negatives, resulting in 83.33% precision, 3.41% recall and 6.56% F-measure, iii) POS features with 1 true positive, 6 false positives, and 292 false negatives, resulting in 14.29% precision, 0.34% recall, and a 0.67% F-measure. Here again, the definiteness feature has failed to produce any positive contributions.

- Named Entity: Symptoms

The results of applying the NBN for the symptoms entity type are shown in Table 5.4. It is only the annotation tags which provide the best results, with 214 true positives, 94 false positives, and 13 false negatives. The precision rate is 69.48%, and the recall rate is 94.27%, while the F-measure is 80.00%. The remaining features have no influence to the classification of this entity. The pattern features record 42 true positives, 7 false positives, and 185 false negatives, giving 85.71% precision, but only 18.50% recall and a total F-measure of 30.43%. The Gazetteer features return 4 true positives, 5 false positives,

and 223 false negatives, giving 44.44% precision, 1.76% recall, and 3.39% total F-measure. The lexical marker features yield 2 true positives, 7 false positives, and 225 false negatives, giving precision of 22.22%, recall of 0.88%, and a F-measure of 1.69%. The poorest results are related to the definiteness features and the POS tags which yield no positives at all for either of them, meaning this analysis could not be completed.

Unlike the disease entity, our gazetteers feature does not yield the most accurate results. This is because our gazetteers contain a limited number of symptoms and because symptoms in this corpus tend to be expressed in terms of sentences rather than a set of lexical items; our corpus includes one symptom entity described in terms of 11 lexical items. The annotation tags appear to be the most influential feature for this kind of medical corpus. Knowing the annotation tags of the previous two or following two words can help determine whether the selected word is a symptom entity. Pattern features have also contributed significantly to the recognition of symptom entities when compared to disease entities, being the fifth most influential feature for the disease entity and the second most influential for the symptoms entity.

Table 5.3 The results of applying the NBN using each feature for the treatment method entity category.

No	Features	TP	FP	FN	Precision	Recall	F-measure
1	Gazetteers features	185	141	108	56.75%	63.14%	59.77%
3	The annotation tags	121	205	172	37.12%	41.30%	39.10%
2	Lexical marker features	14	3	279	82.35%	4.78%	9.03%
5	The pattern features	10	2	283	83.33%	3.41%	6.56%
6	POS tags	1	6	292	14.29%	0.34%	0.67%
4	Definiteness features	0	0	293	N/A	0.00%	N/A

Table 5.4 The results of applying the NBN using each feature for the symptom entity category.

No	Features	TP	FP	FN	Precision	Recall	F-measure
1	The annotation tags	214	94	13	69.48%	94.27%	80.00%
2	The pattern features	42	7	185	85.71%	18.50%	30.43%
3	Gazetteers features	4	5	223	44.44%	1.76%	3.39%
4	Lexical marker features	2	7	225	22.22%	0.88%	1.69%
5	Definiteness features	0	0	227	N/A	0.00%	N/A
6	POS tags	0	0	227	N/A	0.00%	N/A

5.4 Naïve Bayes Network application to named entities

In the previous step, Naïve Bayes Network is used to rank the features according to their performance in recognising the entities. This is done by implementing the NBN using each feature on its own. Contrarily, this section describes the application of NBN using all the features. The structure of our NBN is presented in Figure 5.3. As required by the structure of NBNs, our network consists of one parent node which is the word_i entity tag, which is the target node that needs to be classified. The root node is linked with 27 nodes that represent the features of the previous two words, current word, and the following two words.

A Bayes Server 6.15 tool is used to construct the NBN, to learn the probability of each node and to perform the classification task. A relevance tree algorithm, which is an exact probabilistic inference algorithm for Bayesian networks and dynamic Bayesian networks, is used to perform the learning and inference. Each named entity category is trained and predicted on its own. Consequently, four NBNs were constructed, one for each category.

5.4.1 Evaluation of the NBN classifier

The data was divided into 80% for training and 20% for testing. Table 5.5 below shows the best results obtained in terms of precision, recall and F-measure relate to the entity disease, and the poorest results in terms of F-measure relate to the entity symptoms.

The results show that our NBN achieved the highest F-measure, 79.97%, in recognising the disease entity and the lowest F-measure, 52.47% in recognising the diagnosis methods entity. The results for each named entity are illustrated in Figure 5.4 highlighting the effectiveness of our NBN in predicting (classifying) the entities. The Venn diagram consists of two circles representing the actual entities in the corpus and the predicted entities by our NBN. The intersection, coloured in green, represents the entities which correctly have been predicted (true positives). The pink circle represents the entities which our NBN failed to predict (false negatives) and the blue circle represents the non-entities which have been predicted as entities by our NBN (false positives). It is clearly shown by the Venn diagram that our NBN has labelled a wider range of words as entities, achieving a high recall where a high number of correct entities are detected (large number of true positives and small number of false negatives). However, this has affected the precision (large number of false positive errors). In other words, our NBN has achieved 3.5 false positive errors per false negative error.

Table 5.5 The results from our NBN.

No	Entity	TP	FP	FN	Precision	Recall	F-measure
1	Disease	481	190	51	71.68%	90.41%	79.97%
2	Treatment methods	217	240	76	47.48%	74.06%	57.87%
3	Symptoms	141	151	86	48.29%	62.11%	54.34%
4	Diagnosis methods	207	329	46	38.62%	81.82%	52.47%

5.4.2 Optimising the Naïve Bayes Network performance

It is evident that the use of a high dimensional feature space can lead to a poor classification performance for our NBN. Table 5.6 displays the results of evaluating our NBN using the highest ranked feature on its own (excluding the annotation tag feature), discussed in Section 5.2, against the results of using all the features in one network. It is clear that the performance of the NBN using the complete set of features has achieved a low F-measure score for the disease, treatment method, and diagnosis method entities. However, it has achieved a good F-measure score for the symptoms entity.

Table 5.6 The results of our NBN using two different sets of features: all features and only the highest ranked feature.

No	Entity	All Features	Highest ranked Feature	Difference
1	Diseases	79.97%	85.57%	-5.60%
2	Treatment methods	57.87%	59.77%	-1.90%
3	Symptoms	54.34%	30.43%	+23.91%
4	Diagnosis methods	52.47%	53.97%	-1.50%

To optimise and enhance the performance of our NBN we have reduced the number of nodes representing the features in our NBN. The new optimised network only employs the three highest ranked features in terms of their F-measures for each named entity, within a five-word window. Figure 5.5 illustrates the structure of the optimised NBN which consists of 16 nodes instead of 28 nodes and the results are summarised in Table 5.7.

Table 5.7 The results from the optimised NBN.

No	Entity	TP	FP	FN	Precision	Recall	F-measure
1	Diseases	435	10	97	97.75%	81.77%	89.05%
2	Treatment methods	186	240	107	43.66%	63.48%	51.74%
3	Symptoms	96	122	131	44.04%	42.29%	43.15%
4	Diagnosis methods	136	191	117	41.59%	53.75%	46.90%

Table 5.8 below compares the performance of the NBN based on the number of features used (all features, the highest ranked feature in Section 5.2, and the highest three features in Section 5.2). For the diseases entities, the NBN which is trained with only the three highest ranked features has achieved the highest F-measure score with 89.05%. The NBN trained with only the highest ranked feature has achieved the highest F-measure score for the treatment methods and diagnosis methods entities with 59.77% and 53.97% respectively. The NBN that has been trained with the all features achieved the highest F-measure 54.34% in recognising the symptoms entity. In conclusion, the reduction in the number of features used to train our NBN has improved the performance of NBN in recognising three out of four named entities.

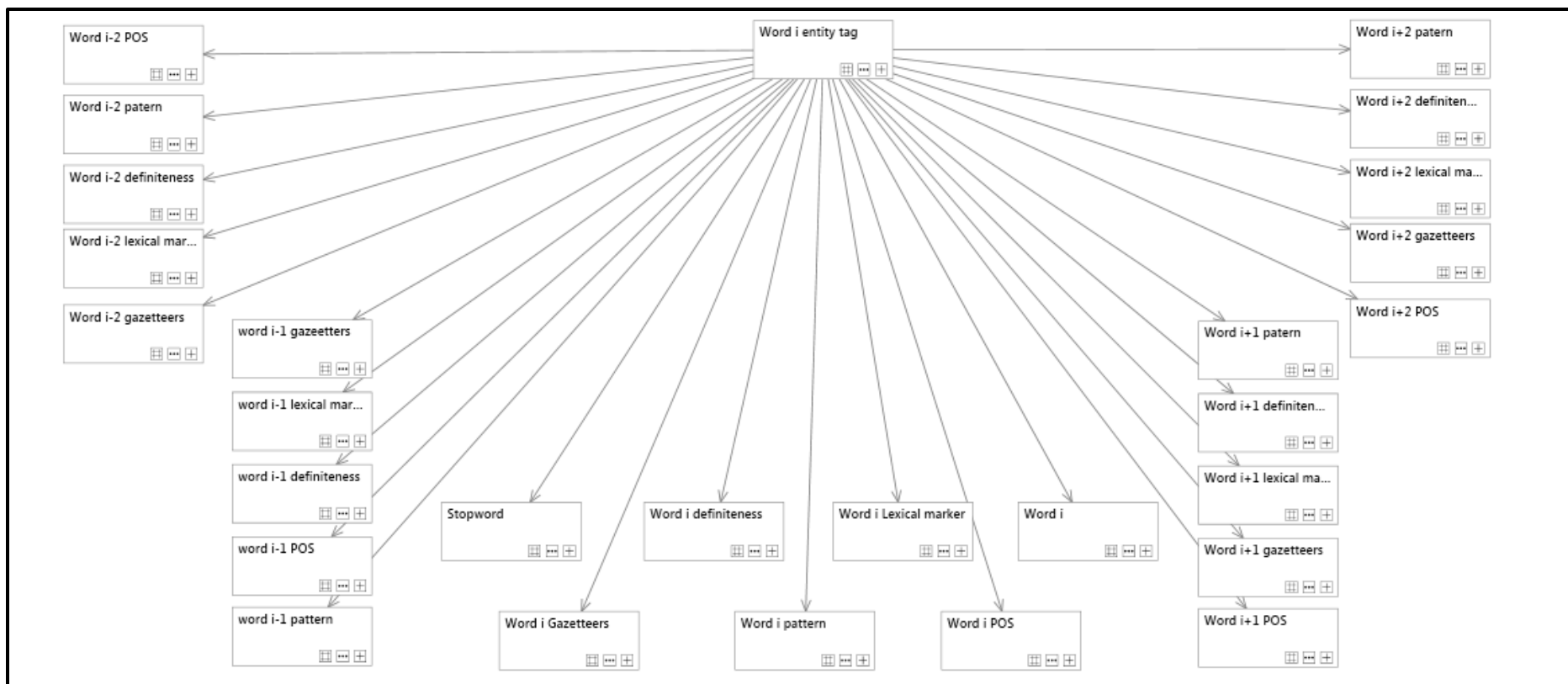


Figure 5.3 The structure of our NBN.

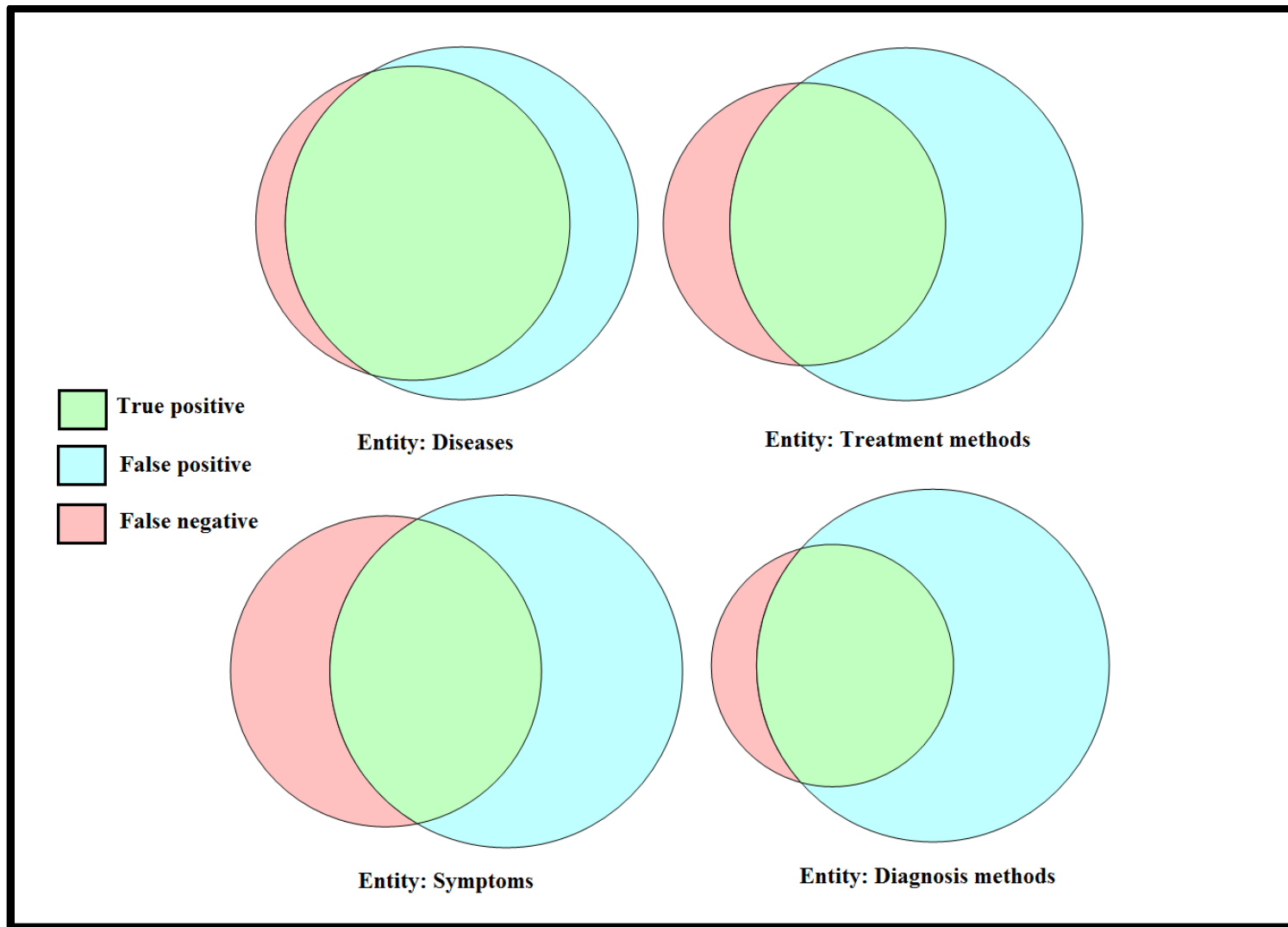


Figure 5.4 The Venn diagram representation of the NBN performance.

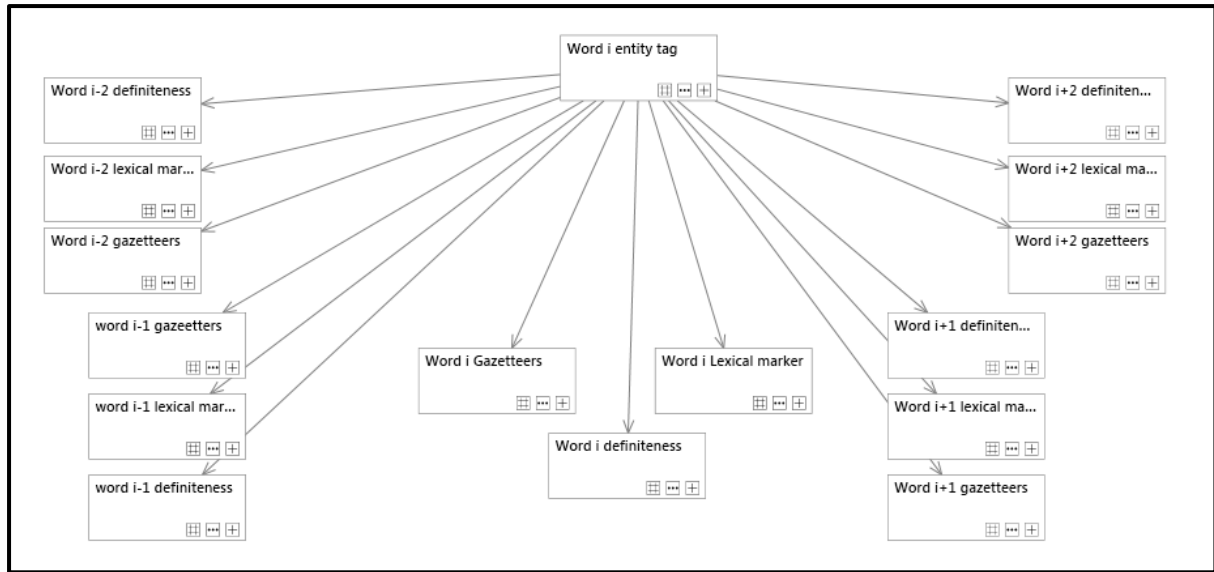


Figure 5.5 The structure of the optimised NBN.

Table 5.8 The performance of the NBN depending on the number of features used.

No	Entity	All Features	The highest ranked feature	The three highest ranked features
1	Diseases	79.97%	85.57%	89.05%
2	Treatment methods	57.87%	59.77%	51.74%
3	Symptoms	54.34%	30.43%	43.15%
4	Diagnosis methods	52.47%	53.97%	46.90%

5.5 Bayesian Belief Network

The performance abilities of NBN can be surprising, given that its main assumption is that all variables in the network are independent, which is an unrealistic assumption in real data. That said, NBN generally delivers fairly accurate classifications. For example, Friedman *et al.* (1997) illustrated this ability with a classifier that assesses the risks present in loan applications. It may appear counterintuitive to ignore the correlations between age, education level, and income to assess each one independently. This sort of scenario is precisely why some researchers contemplate the possibility of improving NBN performance by eliminating the assumption that variables are independent (Friedman *et al.* (1997)).

On the other hand, the Bayesian belief network (BBN) offers greater flexibility when it comes to forming the structure with a classifier. Some flexibility is offered by the absence of restrictions that arise when all the nodes (X_1, X_2, \dots, X_n) need to be assigned to Class C as the child of the parent. More flexibility is found when there can be more than one parent. Due to this flexibility, the relationship between all nodes, including the class nodes, can be incorporated into the structure of the BBN. However, there is the risk that the searching space and the parameter learning can grow exponentially if the number of parents is not monitored and controlled. The BBN improves the classification process through the unrestricted ability to link variables and classes, and the learning structure lends itself to forming a Bayesian network, getting closer to the model required by expert knowledge. Figure 5.7 is an example of a BBN.

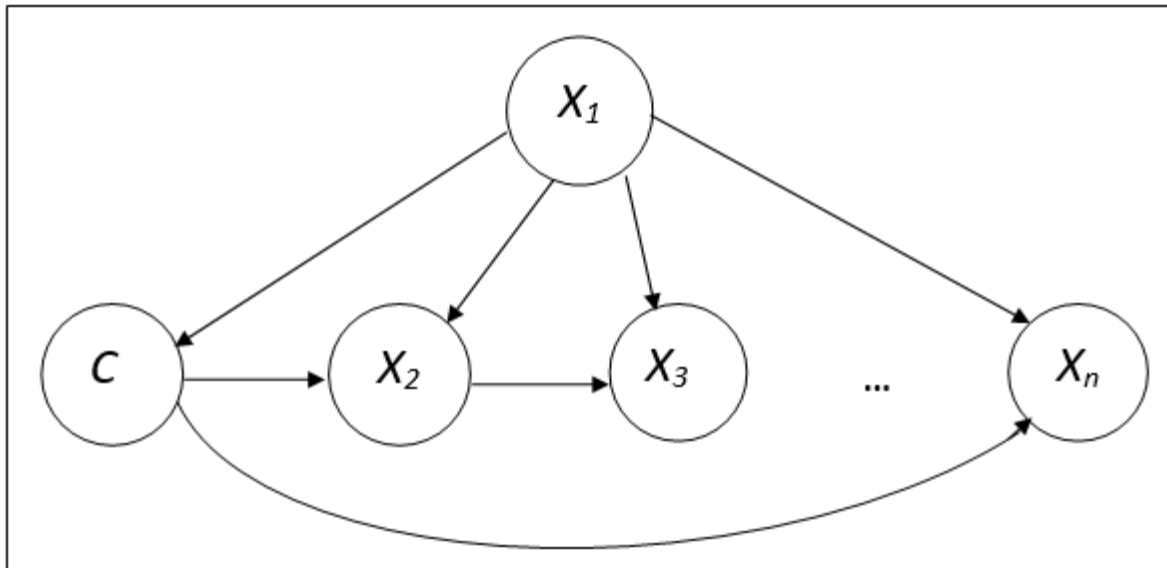


Figure 5.6 An example of a GBN (Ang et al., 2016).

5.5.1 BBN structure

Figure 5.7 shows the structure of our BBN approach to classify the entities. It consists of 31 nodes and 30 arcs. The target node, which needs to be predicted and recognised, is the $word_i$ entity node. The nodes $word_{i-2}$ entity, $word_{i-1}$ entity, $word_{i+1}$ entity and $word_{i+2}$ entity represent the annotation tags of the previous and subsequent two words from the target word. The other nodes represent the values of the lexical markers, gazetteers, and pattern features, definiteness, and POS features for each entity node. Each entity node is a child of three features (lexical marker, gazetteer, and pattern) and a parent

of two features (POS tags and definiteness). The word_{*i*} entity node is also linked with the word_{*i-1*} and the word_{*i+1*} entity nodes.

The algorithm supporting our BBN structure can be summarised as follows:

- Knowing the feature values of the word_{*i*} entity node (e.g., lexical marker, gazetteer, and pattern features, etc.) can lead to better prediction of the value of the word *i* entity node.
- Knowing the entity types of the previous and subsequent two words of the target word (word_{*i-2*}, word_{*i-1*}, word_{*i+1*}, and word_{*i+2*}) can lead to better prediction of the word_{*i*} entity node.
- For a better prediction of the word_{*i-2*}, word_{*i-1*}, word_{*i+1*}, and word_{*i+2*} entity node values, their feature nodes must be considered.

5.4.2 Evaluation of the BBN classifier

In order to carry out the NER task using BBN, the data is again divided into training data, which constitutes 80%, and testing data, which constitutes 20%. During the training phase, values of all nodes are provided to the BBN while during the testing phase, the values of nodes word_{*i*} entity tag, word_{*i-2*} entity tag, word_{*i-1*} entity tag, word_{*i+1*} entity tag, and word_{*i+2*} entity tag are set as missing. A relevance tree algorithm, which is an exact probabilistic inference algorithm for Bayesian networks and dynamic Bayesian networks, is used to perform the learning and the inference. Each named entity is trained and predicted on its own. Table 5.9 lists the results of our BBN network.

Table 5.9 The results of our BBN network.

No	Entity Type	TP	FP	FN	Precision	Recall	F-measure
1	Disease	483	17	49	96.60%	90.79%	93.60%
2	Treatment methods	208	92	85	69.33%	70.99%	70.15%
4	Diagnosis methods	135	24	118	84.91%	53.36%	65.53%
3	Symptom	112	45	115	71.34%	49.34%	58.33%

Our BBN achieved a 96.60% precision, 90.79% recall, and 93.60% F-measure for the disease entity, while for the treatment method entity, it achieved 69.33%, 70.99%, and 70.15% for precision, recall, and F-measure, respectively. For the diagnosis method and symptom categories, our system achieved 84.91% and 71.34%, respectively, for precision, 53.36% and 49.34%, respectively, for recall, and 65.53% and 58.33%, for F-measure, respectively. Figure 5.8 illustrates the effectiveness of our BBN, the green intersection area representing the true positives, the pink area representing the false negatives, and the blue area representing the false positives. Unlike the results of our NBN (see Table 5.6 and Figure 5.4), this BBN structure has significantly decreased the number of false positive errors

as it is shown in Figure 5.9 by improving the labelling of words as entities. This has improved precision but the recall has slightly decreased because the false negative errors have increased. However, the F-measure of this BBN has outperformed the NBN results. Furthermore, the recognition of the diseases entities has achieved the highest precision and recall (the lowest number of errors), shown by the large intersection area. A complete eclipse would occur if there were no errors in classification phase.

Table 5.10 captures the results of the BBN with three different NBN configurations: using all features, only the highest ranked feature and three highest ranked features. The BBN outperforms all the NBNs for all entities in terms of the F-measure with at least 4.55% for the disease entity, 10.38% for the treatment method entity, 11.19% for the symptom entity, and 4.36% for the diagnosis method entity.

Table 5.10 Comparative analysis of the results achieved by the BBN and three NBNs.

No	Entity	NBN (All features)	NBN (the highest ranked feature)	NBN (the three highest ranked features)	BBN (all features)
1	Diseases (D)	79.97%	85.57%	89.05%	93.60%
2	Treatment methods (T)	57.87%	59.77%	51.74%	70.15%
3	Symptoms (S)	54.34%	30.43%	43.15%	65.53%
4	Diagnosis methods (G)	52.47%	53.97%	46.90%	58.33%

5.4.3. Errors analysis

Even though the F-measure is the standard measure by which a NER system is assessed, it does not inform the researcher about the nature of the errors detected. Our experiments would benefit from a detailed analysis of these errors as this would help improve the overall system performance which F-measure statistics alone cannot do. To gain an understanding of the types of errors found within each named entity category, the researcher carried out a manual analysis of the errors for each entity, described below.

- Entity: Diseases

The results show that our BBN had 17 false positives, where non-entities were flagged as entities, and 49 false negatives, where entities were overlooked. Table 5.11 shows the statistics of the errors in the results of our BBN.

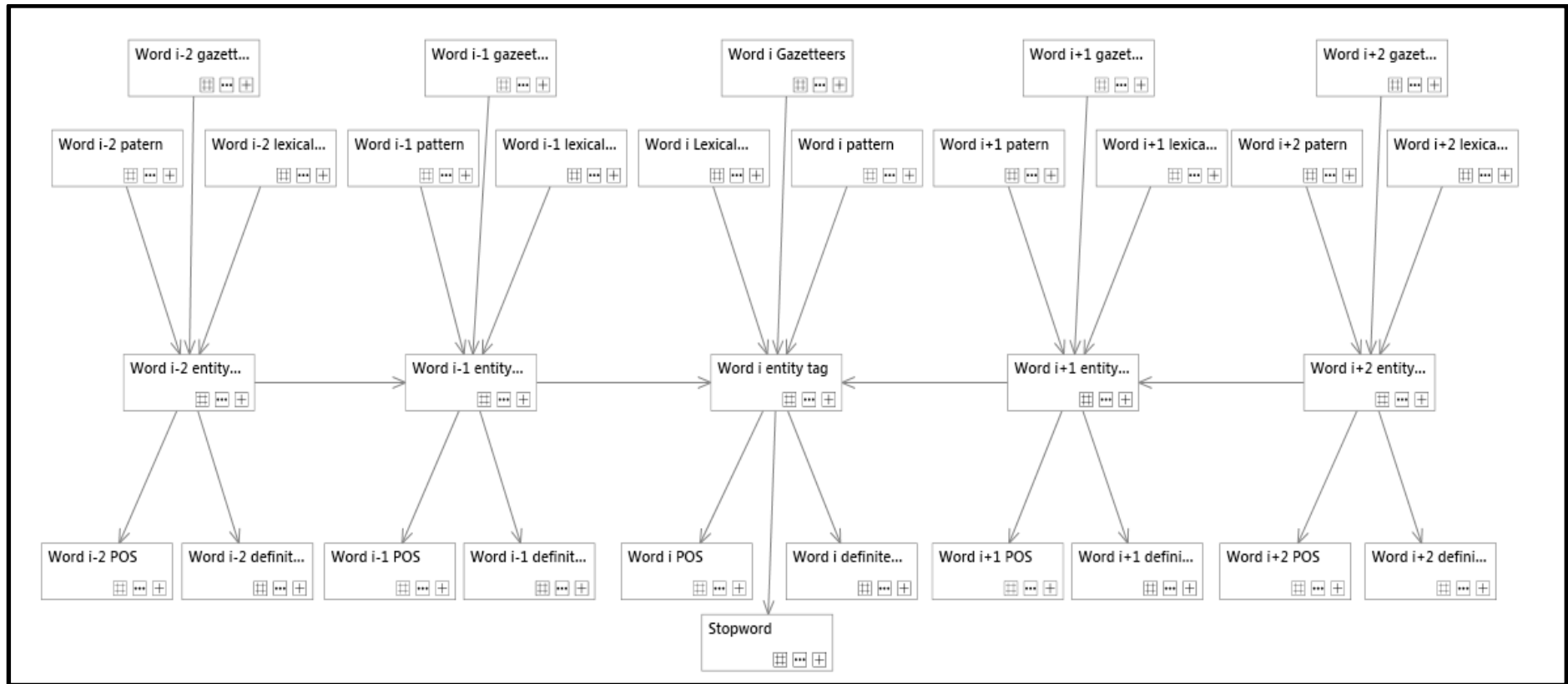


Figure 5.7 The structure of our BBN.

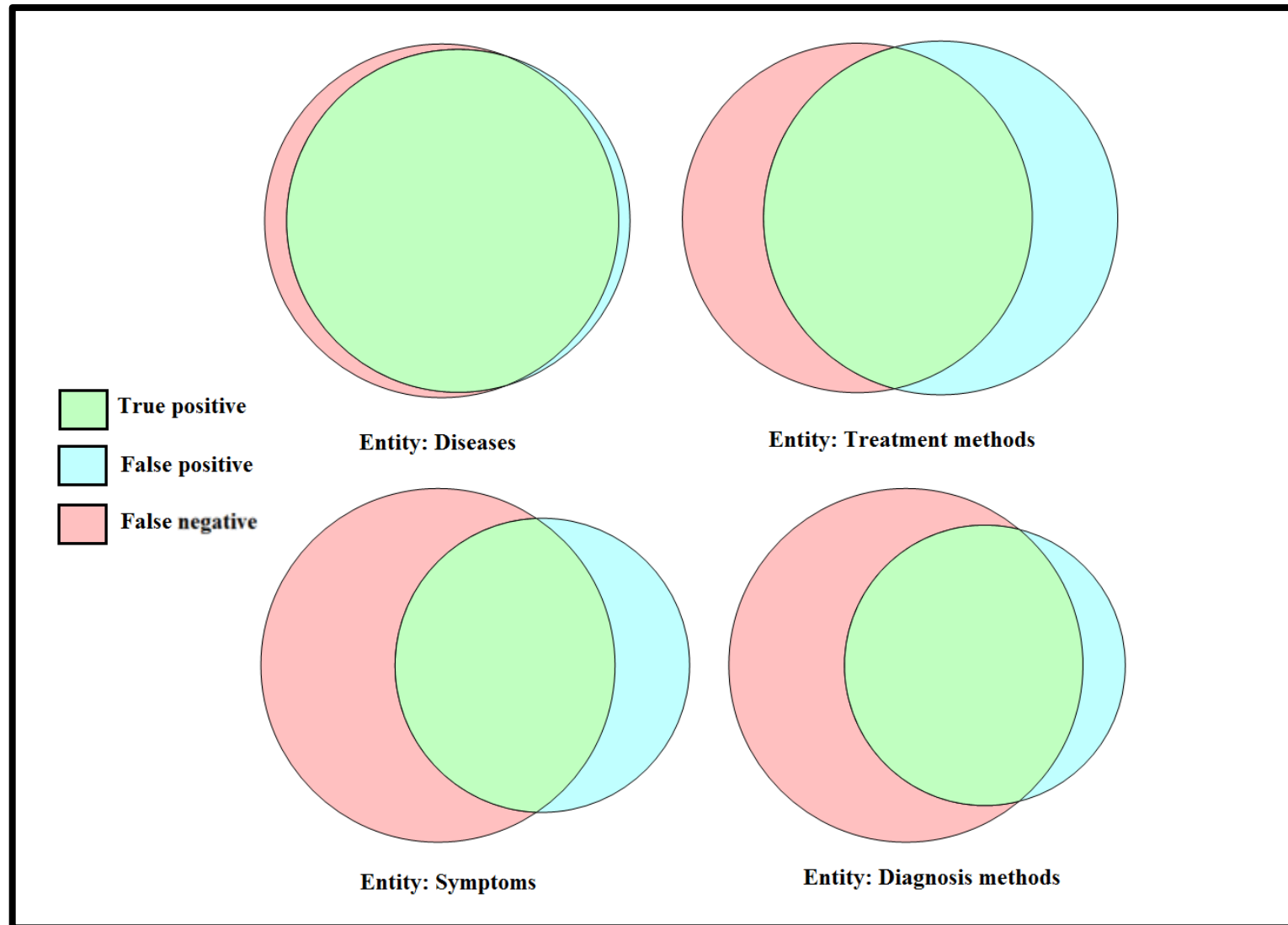


Figure 5.8 The Venn diagram representation of the BBN performance.

Table 5.11 Error results for the disease entity (D).

Actual	Predicted	Count	Probability	Probability Actual	Probability Predicted
Disease	NotEntity	49	0.392%	9.211%	0.408%
NotEntity	Disease	17	0.136%	0.142%	3.400%

Figure 5.9 shows a sample of the output of our BBN. The figure illustrates the original annotation tag (word_i entity tag) of each word (word_i) and the predicted annotation tag and belief probability in the predicted annotation tag (PredictProbability (word_i entity tag)).

Predict(Word i entity tag)	PredictProbability(Word i entity tag)	Word i	Word i entity tag
NotEntity	100.000 %	,	NotEntity
NotEntity	100.000 %	إن	NotEntity
NotEntity	99.323 %	الأمر يكيبين	NotEntity
NotEntity	99.384 %	الأصلين	NotEntity
NotEntity	99.959 %	معرضون	NotEntity
NotEntity	100.000 %	أكثر	NotEntity
NotEntity	100.000 %	من	NotEntity
NotEntity	99.960 %	غيرهم	NotEntity
NotEntity	100.000 %	ل	NotEntity
NotEntity	99.970 %	الإصابة	NotEntity
NotEntity	99.999 %	ب	NotEntity
D	99.907 %	سرطان	D
D	99.902 %	المرارة	D
NotEntity	100.000 %	,	NotEntity
NotEntity	100.000 %	يزداد	NotEntity
NotEntity	98.378 %	خطر	NotEntity
NotEntity	98.481 %	الإصابة	NotEntity
NotEntity	99.999 %	ب	NotEntity
D	99.912 %	سرطان	D
D	99.905 %	المرارة	D
NotEntity	97.507 %	مع	NotEntity
NotEntity	57.565 %	التقدم	NotEntity
NotEntity	100.000 %	في	NotEntity
NotEntity	99.970 %	السن	NotEntity

Figure 5.9 A sample of the output of our BBN when recognising the disease entity.

The error analysis has revealed that most errors are related to the boundary of the named entity, the words preceding the named entity, and the words following the named entity. For example, our BBN labels the word 'مراحل' 'stages', which appears immediately before the named entity in the sentence

‘مراحل سرطان المعدة’ *stomach cancer stages* and the word ‘أعراض’ *symptoms*, which appears before the NE in the sentence ‘أعراض سرطان الرئة’ *the symptoms of lung cancer* as a disease entity. Similarly, our BBN labels the word ‘الجراحة’ *surgery*, which follows the NE ‘سرطان المريء الجراحة’ *Oesophageal cancer treatments include surgery* as a disease entity in the sentence. Another type of error is related to the compound disease names. For instance, in the sentence ‘سرطان الثدي و القولون’ *Breast and colon cancer*, our BBN recognises just the first type of cancer, which is breast cancer and misses the word *colon*. Another common error is labelling the adjective of the word *cancer* as a disease entity. For example, our BBN recognises the sentence ‘سرطان نقلي’ *metastatic cancer* as a disease entity although it is not really a type of cancer but is an adjective describing a cancer which spreads from the place where it first started to another place in the body.

- Entity: Treatment method

The results show that our BBN had 92 false positives, where non-entities were flagged as entities, and 85 false negatives, where entities were overlooked. The number of false positive errors and false negative errors are close to each other. Therefore, the precision of our BBN is quite similar to the recall. Table 5.12 shows the statistics of the errors in the results of our BBN.

Table 5.12 Error results for the treatment method entity (T).

Actual	Predicted	Count	Probability	Probability Actual	Probability Predicted
NotEntity	Treatment	92	0.736%	0.754%	30.667%
Treatment	NotEntity	85	0.680%	29.010%	0.697%

Figure 5.10 shows a sample of the output of our BBN in terms of recognising the diagnosis method entity. The figure illustrates the original annotation tag (*word_i* entity tag) of each word (*word_i*) and the predicted annotation tag and the belief probability in the predicted annotation tag (PredictProbability (*word_i* entity tag)). The error analysis revealed that many of the false positive errors are linked to the word ‘المعالجة’ *treatment* where our BBN considers this word a treatment method entity, even if it is not. This is due to the existence of this word in the gazetteers. In Figure 5.11, the phrase ‘المعالجة المتوفرة’ *the available treatment* is tagged as a treatment method entity although it is not in the annotated data. This clearly shows that gazetteers can mislead the classification model if the entries of the gazetteers are not distinctive. Regarding the false negative errors, some of the treatment method entities tend to appear as a sentence, which makes it a challenging task to recognise the whole entity and the boundaries. For instance, the phrase ‘استئصال الرحم و البوقين و المبيضين التام’ *total abdominal hysterectomy with bilateral salpingo-oophorectomy* is a treatment method entity. This issue is related specifically to NE types among the medical domain and particularly for the symptom entity where one

symptom in our data contained 15 words. Furthermore, it presents a great challenge in medical texts as the entities cannot be easily distinguished by temporal and numeric expressions or capitalisation.

- Entity: Symptoms

The results show that our BBN had 45 false positives, where non-entities were flagged as entities, and 115 false negatives, where entities were overlooked. The precision of our BBN is quite high in comparison with the recall. Thus, our system classification tends to be severe in terms of labelling a

Predict(Word i entity tag)	PredictProbability(Word i entity tag)	Word i	Word i entity tag
NotEntity	100.000 %	و	NotEntity
NotEntity	99.984 %	تشغل	NotEntity
NotEntity	78.214 %	طرق	NotEntity
T	96.544 %	المعالجة	NotEntity
T	98.649 %	المتوفرة	NotEntity
NotEntity	99.777 %	ل	NotEntity
NotEntity	86.279 %	سرطان	NotEntity
T	54.074 %	المرارة	NotEntity
T	73.264 %	الجراحة	T
NotEntity	99.898 %	و	NotEntity
T	99.843 %	المعالجة	T
T	100.000 %	الشعاعية	T
NotEntity	99.895 %	و	NotEntity
T	99.843 %	المعالجة	T
T	100.000 %	الكيميائية	T
NotEntity	81.138 %	عادة	NotEntity
NotEntity	100.000 %	,	NotEntity
NotEntity	100.000 %	أو	NotEntity
NotEntity	100.000 %	أي	NotEntity
NotEntity	99.960 %	مزيج	NotEntity
NotEntity	99.999 %	من	NotEntity
NotEntity	100.000 %	هذه	NotEntity
NotEntity	94.374 %	الطرق	NotEntity
NotEntity	94.201 %	المختلفة	NotEntity

Figure 5.10 A sample of the output of our BBN when recognising the treatment method entity.

word as a symptom entity. Therefore, our system mistakenly labels 45 words and fails to recognise 115 real symptom entities. Table 5.13 shows a statistical analysis of the errors generated by our BBN.

Symptoms are usually expressed in terms of long syntactic phrases and can reach up to 15 tokens for one entity. For instance, one of the symptoms in our corpus is ‘وجود كتلة في منطقة الأذن أو الوجنة أو الفك أو’ ‘*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. The error analysis has shown that our BBN could recognise symptoms of up to five tokens correctly. Figure 5.11 shows a sample of the output of our BBN in terms of recognising the diagnosis method entity. The figure illustrates the original annotation tag (word_i entity tag) of each word (word_i) and the predicted annotation tag and belief probability in the predicted annotation tag (PredictProbability (word_i entity tag)). Unlike other entities, symptom entities could have stop words among them. For instance, (Figure 5.11), the words في in the sentence ‘الشعور بالثقل في الحوض’ ‘*heavy feeling in the pelvis*’ and the word من in the sentence ‘نزف من المهبل’ ‘*bleeding from the vagina*’ are stop words. Generally, most stop words are not entities; therefore, our BBN did not recognise the previous words as a part of the symptom entity.

Table 5.13 Error results for the symptom entity (S).

Actual	Predicted	Count	Probability	Probability Actual	Probability Predicted
Symptoms	NotEntity	115	0.920%	50.661%	0.932%
NotEntity	Symptoms	45	0.360%	0.367%	28.662%

- Entity: Diagnosis methods

The results show that our BBN had 24 false positives, where non-entities were flagged as entities, and 118 false negatives, where entities were overlooked. Similarly, to the symptom entity results, the precision of our BBN is quite high in comparison with the recall. Thus, our system classification tends to be coarse in terms of labelling a word as a diagnosis method entity. Therefore, our system mistakenly labels 24 words and fails to recognise 118 real diagnosis method entities (Table 5.14).

Table 5.14 Error results for the diagnosis method entity (G).

Actual	Predicted	Count	Probability	Probability Actual	Probability Predicted
Diagnosis methods	NotEntity	118	0.944%	46.640%	0.956%
NotEntity	Diagnosis methods	24	0.192%	0.196%	15.094%

The error analysis has revealed that most of the errors are related to either the boundary of the NEs or the failure of detecting the whole NE. Figure 5.12 shows a sample of the output of our BBN in terms of recognising the diagnosis method entity. The figure illustrates the original annotation tag (word i entity tag) of each word (word i) and the predicted annotation tag and belief probability in the predicted annotation tag (PredictProbability (word i entity tag)). By considering the examples in Figure 5.12, our BBN failed to detect the beginning of the diagnosis method entity 'تصوير الصدر بالأشعة السينية' 'chest imaging with X-ray'. The term X-ray is a diagnosis method entity; however, the whole sentence was tagged as a diagnosis method entity during the annotation stage. Some of the false negative errors are repeated many times among our corpus. Hence, the recall and overall performance of our BBN dropped.

Predict(Word i entity tag)	PredictProbability(Word i entity tag)	Word i	Word i entity tag
NotEntity	99.848 %	و	NotEntity
NotEntity	100.000 %	قد	NotEntity
NotEntity	99.641 %	تتضمن	NotEntity
NotEntity	99.460 %	الأعراض	NotEntity
NotEntity	99.779 %	ما	NotEntity
NotEntity	98.075 %	يلي	NotEntity
NotEntity	100.000 %	:	NotEntity
S	68.367 %	الشعور	S
S	82.291 %	ب	S
S	68.368 %	الثقل	S
NotEntity	75.649 %	في	S
S	73.017 %	الحوض	S
NotEntity	100.000 %	.	NotEntity
S	85.089 %	الألم	S
NotEntity	65.925 %	أسفل	S
S	69.017 %	البطن	S
NotEntity	100.000 %	.	NotEntity
S	86.497 %	نزف	S
NotEntity	58.460 %	من	S
S	72.472 %	المهبل	S
NotEntity	100.000 %	.	NotEntity

Figure 5.11 A sample of the output of our BBN when recognising the symptom entity.

5.5 The Effect of the Sliding Window Size on the BBN Performance

Normally the neighbouring words to the left and right of the target word can convey important information, which will assist in identifying named entities. Sliding windows allow the researcher to factor in some of the context when a classification decision is made for any given word. By including the features of the preceding and following words as well as those of the targeted word, the decision regarding the targeted word acknowledges the semantic value of its features as well as the features of other words in the window. In order to measure the effect of the sliding window size on the performance of our BBN, seven different sliding window sizes were examined for each NE category. These sizes are as follows:

Predict(Word i	PredictProbability(Word i entity tag)	Word i	Word i entity tag
NotEntity	100.000 %	أن	NotEntity
NotEntity	99.998 %	يكون	NotEntity
NotEntity	98.991 %	تصوير	G
NotEntity	88.137 %	الصدر	G
NotEntity	92.765 %	ب	G
G	99.685 %	الأشعة	G
G	99.524 %	السينية	G
NotEntity	93.618 %	و	NotEntity
G	73.501 %	التصوير	G
G	100.000 %	المقطعي	G
G	99.925 %	المحوري	G
NotEntity	98.432 %	و	NotEntity
NotEntity	92.985 %	ب	NotEntity
G	100.000 %	الرئين	G
G	100.000 %	العفناطيسي	G
G	62.907 %	مفيدا	NotEntity
NotEntity	99.931 %	في	NotEntity
NotEntity	99.022 %	تحديد	NotEntity
NotEntity	98.625 %	المرحلة	NotEntity
NotEntity	100.000 %	التي	NotEntity
NotEntity	99.911 %	بها	NotEntity
NotEntity	99.374 %	السرطان	NotEntity
NotEntity	99.997 %	.	NotEntity

Figure 5.12 A sample of the output of our BBN when recognising the diagnosis method entity.

- **-/+Two-word window:** The features of the preceding and following two words in addition to the features of the current word.
- **-/+One-word window:** The features of the preceding and following word in addition to the features of the current word.
- **One-word window:** The features of the current word only.
- **-Two-word window:** The features of the preceding two words in addition to the features of the current word.
- **-One-word window:** The features of the preceding word in addition to the features of the current word.
- **+Two-word window:** The features of the following two words in addition to the features of the current word.
- **+One-word window:** The features of the following word in addition to the features of the current word.

The sizes illustrate and visualise the different sliding window sizes given the following sentence ‘Lung cancer is one of the most common cancers in the world’ where the target word is *of* (Figure 5.13).

-/+2 words window	Lung	cancer	is	one	of	the	most	common	cancers
-/+1 word window	Lung	cancer	is	one	of	the	most	common	cancers
1 word window	Lung	cancer	is	one	of	the	most	common	cancers
-2 words window	Lung	cancer	is	one	of	the	most	common	cancers
-1 words window	Lung	cancer	is	one	of	the	most	common	cancers
+2 words window	Lung	cancer	is	one	of	the	most	common	cancers
+1 words window	Lung	cancer	is	one	of	the	most	common	cancers

Figure 5.13 Visualisation of different sliding window sizes.

- Entity: Diseases

In order to carry out the experiments, our BBN size is repeatedly adjusted according the sliding window size. Data are divided into 80% training data and 20% testing data. Table 5.15 lists the results of our BBN network per sliding window size. Our BBN with a $-/+$ one-word window achieved the highest recall at 90.98% and the highest F-measure at 94.07%. The one-word window achieved the highest precision with 98.46%. However, it achieved a quite low recall at only 47.93% and its F-measure was only 64.48%. The low recall is caused by ignoring the useful information that relies on the contextual words. It is worth noting that the $-$ one-word and $-$ two-word windows achieved high F-measures in comparison with the $+$ one- and $+$ two-word windows. This might reflect the nature of the Arabic language, where usually the keywords appear before the entity. For instance, in the phrase ‘سرطان الرئة’ ‘lung cancer’, unlike English, the word *cancer* precedes the word *lung*. Therefore, the phrase is written and pronounced as ‘*cancer lung*’. Thus, considering the features of the previous one or two words will result in a better performance than when considering the following one or two words features.

Table 5.15 The results of our BBN network per sliding window size for the disease entity.

No	Sliding window	TP	FP	FN	Precision	Recall	F-measure
1	$-/+$ 2-word window	483	17	49	96.60%	90.79%	93.60%
2	$-/+$ 1-word window	484	13	48	97.38%	90.98%	94.07%
3	1-word window	255	4	277	98.46%	47.93%	64.48%
4	$-$ 2-word window	482	16	50	96.79%	90.60%	93.59%
5	$-$ 1-word window	438	8	94	98.21%	82.33%	89.57%
6	$+$ 2-word window	299	80	233	78.89%	56.20%	65.64%
7	$+$ 1-word window	346	75	186	82.19%	65.04%	72.61%

- Entity: Diagnosis methods

Table 5.16 shows the results of applying the different sliding window sizes using our BBN on the diagnosis method entity. Our BBN with a one-word window achieved the highest precision at 88.68%. However, its recall was the lowest of all BBNs at 37.15%, and its F-measure was also the lowest at 52.37%. Again, ignoring the surrounding words and the contextually relevant information they convey resulted in a poor recall percentage. On the other hand, our BBN with a $-/+$ two-word window achieved the highest recall and F-measure at 53.36% and 65.53%, respectively, although its precision was only recorded at 84.91%. Here the $-$ two-word and $-$ one-word windows scored lower on

precision, recall, and F-measure than the +two-word and +one-word windows, suggesting that, here, the adjective and noun positioning is not as relevant as it was for the disease entity category. None of the BBNs scored over 53.6% for recall or over 65.53% for F-measure, which is significantly lower than for the disease entity category. This may be due to the complexity of most diagnosis method entities.

Table 5.16 The results of our BBN network per sliding window size for the diagnosis method entity.

No	Sliding window	TP	FP	FN	Precision	Recall	F-measure
1	-/+ 2-word window	135	24	118	84.91%	53.36%	65.53%
2	-/+1-word window	134	23	119	85.35%	52.96%	65.37%
3	1-word window	94	12	159	88.68%	37.15%	52.37%
4	-2-word window	119	33	134	78.29%	47.04%	58.77%
5	-1-word window	119	30	134	79.87%	47.04%	59.20%
6	+2-word window	122	28	131	81.33%	48.22%	60.55%
7	+1-word window	122	18	131	87.14%	48.22%	62.09%

- Entity: Treatment methods

Table 5.17 shows the results of applying the different sliding window sizes using our BBN on the treatment method entity. Concerning the treatment method entity category, there is a less variability in terms of precision. The window with the highest precision percentage is the -one-word window at 73.31%, which also has the highest F-measure at 71.78%. However, the -/+two-word window has the highest recall at 70.99%. Nonetheless, unlike the disease entity and diagnosis method categories, there is not much variation in the results of all the word windows in terms of precision, recall, or F-measure. Precision varies from 69.33% to 73.31%, recall from 61.09% to 70.99%, and F-measure from 65.82% to 71.78%. This suggests that the context of the treatment method entity is less significant than the context of a disease entity or diagnosis method entity. However, the “-1-word” and “-2-word” windows achieved high F-measures in comparison with the “+1-word” and “+2o-word” windows. Thus, context which relies in the previous words of the target words is more significant than the context of the following words.

Table 5.17 The results of our BBN network per sliding window size for the treatment methods entity.

No	Sliding window	TP	FP	FN	Precision	Recall	F-measure
1	-/+ 2-word window	208	92	85	69.33%	70.99%	70.15%
2	-/+1-word window	207	89	86	69.93%	70.65%	70.29%
3	1-word window	179	58	114	75.53%	61.09%	67.55%
4	-2-word window	207	78	86	72.63%	70.65%	71.63%
5	-1-word window	206	75	87	73.31%	70.31%	71.78%
6	+2-word window	181	76	112	70.43%	61.77%	65.82%
7	+1-word window	181	72	112	71.54%	61.77%	66.30%

- Entity: Symptoms

Table 5.18 shows the results of applying the different sliding window sizes using our BBN on the symptom entity. For the symptom entity category, the most accurate BBN across the board was the -/+two-word window at 71.34% precision, 49.34% recall, and 58.33% F-measure. Furthermore, each measure of accuracy is highly variable across the BBN windows, with precision measuring between 51.94% and 71.34%, recall between 25.99% and 49.34%, and F-measure between 36.53% and 58.33%. These two observations suggest that the context is highly significant in categorising NEs in the symptom entity category. This might be because symptoms are usually sentences with a much larger token count than other entities. This illustrates the need for a larger window size. However, these are still the lowest accuracy percentages of all BBN windows in all four categories. It will be necessary to consider in the future the reasons for this and to adjust future experiments to ensure greater accuracy in the symptom entity category.

Table 5.18 The results of our BBN network per sliding window size for the symptom entity.

No	Sliding Window	TP	FP	FN	Precision	Recall	F-measure
1	-/+ 2-word window	112	45	115	71.34%	49.34%	58.33%
2	-/+1-word window	87	48	140	64.44%	38.33%	48.07%
3	1-word window	59	37	168	61.46%	25.99%	36.53%
4	-2-word window	63	39	164	61.76%	27.75%	38.30%
5	-1-word window	73	66	154	52.52%	32.16%	39.89%
6	+2-word window	67	62	160	51.94%	29.52%	37.64%
7	+1-word window	64	47	163	57.66%	28.19%	37.87%

5.6. Five-fold Cross-validation

The k -fold cross-validation is usually used with a scoring method to avoid over-fitting; the data set can be randomly divided into k -folds of equal size. Each fold is employed as a testing set, and the remaining folds are then applied as a training set, and the test results are averaged over the rounds. The same split must be replicated for training and testing, so that when comparing evaluation results the precision and recall values remain accurate (Benajiba *et al.*, 2010). This method has advantages and disadvantages compared to other methods. On the upside, all observations are used equally for training and validation, with each observation being used exactly once for validation. However, the downside is that the training algorithm has to be rerun k times from scratch, which means it will take k times as much computation to complete an evaluation. The researcher made use of this method, employing a five-fold cross-validation experiment for each named entity category.

- Entity: Disease

Table 5.19 shows the results of five rounds of the experiment for the disease entity. The overall result for the disease entity is 89.31%, 86.21%, and 87.73 for precision, recall, and F-measure, respectively. The highest precision was achieved in Round 1 at 96.60%, while the lowest precision was in Round 3 at 84.35%. The highest obtained recall was achieved in Round 2 at 96.49%, while the lowest recall was in Round 4 at 74.80%. In terms of the F-measure, the highest result was achieved in Round 1 at 93.60%, and the lowest result was in Round 4. The F-measures varied from the average result of the system according to the round at 5.87% up or 4.48% down.

Table 5.19 Disease entity: five rounds experiment.

Round	Precision	Recall	F-measure
One	96.60%	90.79%	93.60%
Two	84.49%	96.49%	90.09%
Three	84.35%	87.94%	86.11%
Four	93.84%	74.80%	83.25%
Five	89.26%	82.94%	85.98%
Overall	89.31%	86.21%	87.73%

- Entity: Diagnosis methods

Table 5.20 shows the results of five rounds of the experts for the diagnosis method entity. The overall result is 83.62% for precision, 47.21% for recall, and 60.35% for F-measure. The highest precision

was in Round 2 at 93.28%, and the lowest was in Round 4 at 72.64%. The highest recall was in Round 1 at 53.36%, and the lowest was in Round 5 at 39.80%. The highest F-measure was in Round 1 at 65.53%, and the lowest was in Round 5 at 52.00%. The F-measures hovered around the overall average; however, the precision and recall ratings displayed greater variability.

Table 5.20 Diagnosis method entity: five rounds experiment.

Round	Precision	Recall	F-measure
One	84.91%	53.36%	65.53%
Two	93.28%	44.94%	60.66%
Three	88.05%	47.30%	61.54%
Four	72.64%	50.00%	59.23%
Five	75.00%	39.80%	52.00%
Total	83.62%	47.21%	60.35%

- Entity: Treatment methods

Table 5.21 shows the results of five rounds of the experiments for the treatment method entity. The overall result is 67.03% for precision, 69.63% for recall, and 68.31% for the F-measure. The highest scores for all three were in Round 3, and the lowest scores for all three were in Round 5. The highest precision was 71.05%, and the lowest was 63.37%. The highest recall was 89.57%, and the lowest was 57.86%. The highest F-measure was 79.25%, and the lowest was 60.49%.

Table 5.21 Treatment method entity: five rounds experiment.

Round	Precision	Recall	F-measure
One	69.33%	70.99%	70.15%
Two	65.10%	64.59%	64.84%
Three	71.05%	89.57%	79.25%
Four	66.06%	70.43%	68.17%
Five	63.37%	57.86%	60.49%
Total	67.03%	69.63%	68.31%

- Entity: Symptom

Table 5.22 shows the results of five rounds of the experiment for the symptom entity. The overall result is 64.39% for precision, 44.85% for recall, and 52.87% for the F-measure, which were the lowest results across all four NE categories. The highest scores for all three measures were in Round 1. The highest score for precision was 71.34%, and the lowest score was in Round 3 at 53.80%. The

highest score for recall was 49.34%, and the lowest score was in Round 5 at 41.06%. The highest score for the F-measure was 58.33%, and the lowest score was also in Round 5 at 46.87%. Both the highest and lowest results as well as the overall results represent the lowest scores of their type in all four categories.

Table 5.22 Symptom entity: five rounds experiment.

Round	Precision	Recall	F-measure
One	71.34%	49.34%	58.33%
Two	71.04%	48.91%	57.93%
Three	53.80%	47.49%	50.45%
Four	70.25%	38.41%	49.66%
Five	54.59%	41.06%	46.87%
Total	64.39%	44.85%	52.87%

5.7 Summary

This chapter has described the steps involved in the second stage of our system which is the classification and named entity recognition stage. The main findings of this stage are summarised below.

- The performance of any machine learning based NER system is highly correlated with the features used to train it. As Table 5.8 has shown, the performance of the NBNs has varied depending on the number of features used. Therefore, ranking the features according to their performance and selecting the optimal feature set would significantly improve the performance of the NER system.
- Despite the simplicity of naïve Bayes, it can deliver fairly accurate classifications for NER purposes, especially with a small number of features (Section 5.3 and 5.4).
- For NER purposes, our Bayesian belief network outperforms the naïve Bayes in terms of the F-measure with 7.5%.
- Minimising the number of parameters within the BBN is preferred, as the parameters number can dramatically increase if one node has many parents and then the classification process could be delayed or even stopped. The minimisation could be done by switching links to the other way round or by avoiding linking nodes with large discrete variables to another node with a large discrete variable.
- The issue regarding the large number of parameters does not exist for naïve Bayes because the structure of naïve Bayes allows only one parent for each node. Thus, naïve Bayes network is the most common form of Bayesian network used for classification tasks. However, naïve Bayes network did not deliver the best classification results in this study.

- The large size of the sliding window could affect the performance of BBN classifier as the -/+one-word window outperformed the -/+two-word window.
- For Arabic NER purposes, the contextual information that relies on the words that come before the target word has led to a better prediction of the entities than the contextual information of the words that come after the target word.
- The entities in our medical domain are often expressed in a complex set of tokens and are not represented by temporal or capitalisation markers; this has made the recognition and learning tasks in our approach very challenging.

Chapter 6: Conclusion

6.1 Review of the study

The aim of this study is to investigate the BBN approach to recognise and extract cancer related named entities from Arabic medical corpus. To this end, a system referred to as NAMERAMA in this thesis is developed. It consists of two main stages. The first stage, which relates to Arabic language processing, comprises three main steps: pre-processing, data analysis, and feature extraction, whereas the second stage focuses on the application of the BBN to classify, recognise, and extract the relevant named entities. Our pre-processing method has included data tokenisation, POS-tagging, and data annotation. Frequency, collocation and concordance analysis have been used to analyse the corpus and to extract the optimal features set. This has enabled us to prepare our corpus adequately to extract and transform relevant features and to be analysed using BBN software.

In summary, the objectives outlined in Chapter 1 have been met as follows:

Objective 1, which refers to the survey of Arabic NER systems and methodologies, is provided in Chapter 2. Although Arabic NER has made some progress, it continues to lag behind NER tools and techniques for similarly important languages. Furthermore, modern Arabic texts have received less research attention than English documents. Due to the complex and varied morphology of the Arabic language structure, it has become necessary to continue the development of toolkits and software systems, corpora, and methods to support future text processing research into the Arabic language.

Objective 2 relates to the much-needed corpus to support research into Arabic NER. The selected and annotated corpus for this study focused on the medical corpus, which was extracted from the King Abdullah Bin Abdullahziz Arabic Health Encyclopedia. This corpus has provided the study with the four classes of named entities namely disease, diagnosis, treatment, and symptoms. This new corpus is a valuable addition to the currently available corpora supporting further research into Arabic language processing.

Objective 3 is concerned with the development of a novel BBN approach to NER. The literature review of NER has identified two important needs, covered in Chapter 3. The first need is to identify the most suitable natural language processing approach to analyse the corpus in order to extract the above classes of named entities. The traditional approach of language processing of English texts has been also useful and applicable to the analysis of Arabic documents. Other linguistic tasks, such collocation and concordance, have also played an important role in identifying named entities. The second need is to investigate how BBN can be used to learn from the annotated data and its linguistic features to predict and recognise the named entities.

Objective 4 is related to the development of our system, NAMERAMA, which is described in two chapters. Chapter 4 focused on the linguistic analysis of the corpus whereas chapter 5 was concerned with the application of BBN to predict and extract the named entities. As this approach was new to the extraction of named entities from Arabic documents, a number of challenges had to be overcome. The most significant challenge was related to the agglutinative nature of Arabic; it became important to ensure that accurate tokenisation is achieved and made effective use of contextual information. Another challenge was related to the configuration of the Bayesian networks as there is no standardised approach and it tends to be domain and goal specific. Two belief networks were applied; the experimental study showed that the belief network has outperformed naïve Bayes with a 7.5% F-measure. The study has also identified the need to minimise the number of parameters in the belief network. As our corpus is extracted from a medical domain this has presented its own challenge; some entity classes were expressed in terms of complex sets of tokens without any temporal or classification markers, making recognition and learning tasks much more complex than the research projects described in chapter 2.

Objective 5, which is related to the evaluation of our approach, is also included in chapter 5. A 5-fold cross validation was applied to validate the developed system. To the best of our knowledge our literature review has not revealed any study from modern standard Arabic medical domain; thus, comparing our system results to other results is difficult. However, there are several Arabic NER systems that recognise different sets of entities such as persons, organisations, and locations, but comparing our results to theirs is not significant because our system uses different data and extracts different sets of entities. Instead, this study has implemented a baseline system to recognise the same entities to compare our system with a baseline system. Our baseline system is based on gazetteers; therefore, it automatically labels a token with an appropriate entity tag whenever this token appears within a gazetteer file.

Table 6.1 shows that the baseline system slightly outperforms our BBN approach in terms of precision for the disease, treatment method, and diagnosis method entities, but not for the symptom entity. On the other hand, our system significantly outperforms the baseline system in terms of the recall for all entities. This result shows that the baseline system achieved high precision with fewer false positive errors, but this leads to a poor recall. There is usually a trade-off between recall and precision. Intuitively, if the system labels a wider range of words as entities, it will detect more correct entities (recall), but it will also get more false errors (lower precision). If it classifies everything in the positive category, it will have 100% recall and bad precision and generally will not be useful as classifier. Thus, the F-measure is a harmonic mean that gives equal weight to recall and precision. In terms of the F-measure, our BBN system outperforms the baseline system by 31.74% for the disease

entity, 3.10% for the treatment method entity, 16.71% for the diagnosis method entity, and 28.68% for the symptom entity (Figure 6.1)

Although the F-measure highlights the performance contribution of our BBN approach, the recall and precision of the four entities, and in particular the treatment methods and symptom entities, require further future work. The complexity of the Arabic language and its computational processing have posed a few limitations which are discussed in Section 6.2.

Table 6.1 The results of the BBN and baseline systems.

No	System Entity	BBN system			Baseline system		
		Precision	Recall	F-measure	Precision	Recall	F-measure
1	Disease	96.60%	90.79%	93.60%	98.36%	45.11%	61.86%
2	Treatment methods	69.33%	70.99%	70.15%	75.86%	60.07%	67.05%
4	Diagnosis methods	84.91%	53.36%	65.53%	95.40%	32.81%	48.82%
3	Symptom	71.34%	49.34%	58.33%	55.42%	20.26%	29.68%

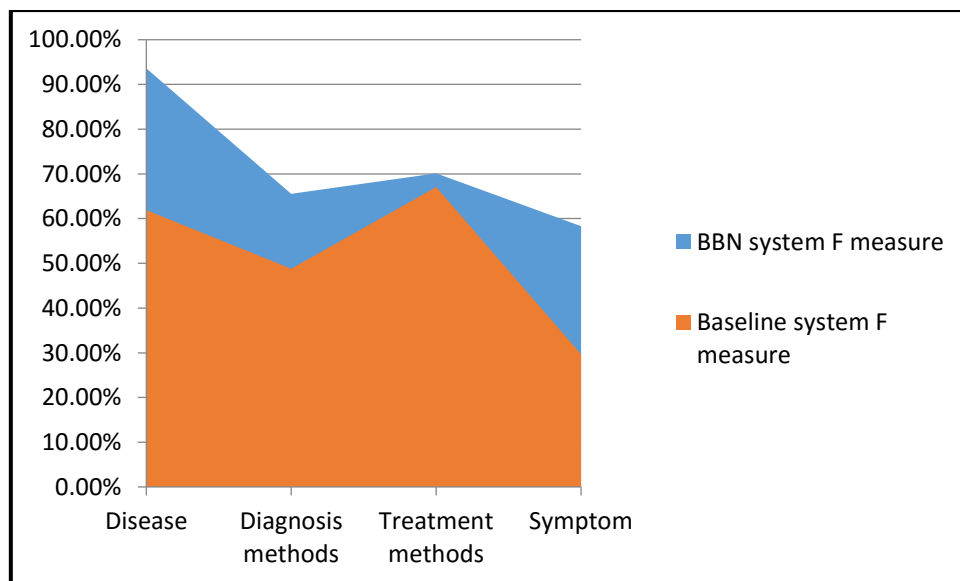


Figure 6.1 F-measure of BBN and baseline system.

6.2 Complexity of Arabic Language and Limitations

As was discussed in Section 1.7, the Arabic language presents many distinctive challenges which are not yet fully addressed by the Natural Language Processing community. However, the domain of medical texts also presents additional challenges, which, combined with Arabic's challenges, may make computational processing a daunting task. A summary of these challenges and their limitations are described below.

- The extraction of entities

In our research we have focused on extracting named entities specific to the medical domain; these entities have not previously been explored by the Arabic NER research community. Most NER systems in the literature have aimed at recognising people's names, organisations, and locations which are easier to detect because they are usually represented by a smaller number of tokens. Unlike aforementioned entities, entities in the medical domain tend to be expressed in terms of a larger set of tokens which makes it difficult to detect the boundaries of multi-word entities in Arabic, due to its complex morphology and different syntax structure (Alotaibi, 2015). Althobitiy (2016) noted that named entities that are only one or two words long constitute 92.09% out of the total named entities in the well-known ANERcorp corpus and constitute 84.39% in her own corpus (ALQAMAR). Furthermore, for the NewsFANE_{Gold}, WikiFANE_{Gold}, and WikiFANE_{Auto} corpora, the percentages are 88.96%, 83.30%, and 86.22, respectively (Alotaibi, 2015). In our corpus, the percentage of named entities that consist of one or two words are 81% for disease entity, 85% for treatment method entity, 66% for the diagnosis method entity, and only 41% for the symptom entity. Moreover, Alotaibi (2015) observed that the error rate in his study was only 39% for single-word named entities while, for multi-words, the error rate was increased to 53%. In the same manner, our system performed 20% better for named entities represented by fewer words (disease and treatment method entities) than named entities represented by more than two words (symptoms and diagnosis methods entities). This brings to light the challenges encountered in extracting such entities in the medical domain.

- Ambiguity

Arabic is an ambiguous language and hence it is difficult to process it automatically (Karov and Edelman, 1998). As a result of this, much research effort was devoted to developing systems for the disambiguation of words in Arabic. It is worth pointing out that the challenge of disambiguation of Arabic words is part of our NER system as many of our named entities are noun based. For instance, the word “سرطان” could mean the disease *cancer* or the animal *crab*, the word “ألم” could mean *pain* or the question *haven't you*, and the word “نقص” could be a noun that means *loss* in the phrase *weight loss*, could be a verb that means *tell a story*, or a verb that means *cut*.

- Agglutination

Arabic is an agglutinative language. An Arabic word may consist of prefixes, a stem or a root, and sometimes even more than one, as well as suffixes with different combinations. Due to this, one Arabic word may be expressed in a sentence in other languages like English. For example, the word وسيتذكروننا is equivalent to the sentence: “and they will remember us”. To overcome this challenge and as a part of the pre-processing steps, AMIRA tool has been used to tokenise our corpus achieving 91.30%, 88.53%, and 89.90% for precision, recall, and F-measure, respectively. Then the errors were corrected manually by the researcher.

6.3 Novel Contributions

This study has achieved a set of novel and valuable contributions to NER and Arabic NLP. These contributions, which can contribute to the advancement of research into both Arabic language processing and medical NER tasks, are outlined below.

- The application of BBN in the context of NER task

To the best of our knowledge, no previous study has implemented BBNs to extract and recognise NEs from Arabic texts. Roth and Yih (2002) have performed similar research in English, however we have found no such research in Arabic. Furthermore, Roth and Yih's work was very limited in terms of the corpus employed, to fewer than a thousand sentences in total focused on four NEs. Furthermore, their system can only recognise entities in sentences that contain two specific verbs (kill - born in). This research has covered complex set of NEs expressed in terms of many tokens extracted from a larger corpus. The developed BBN approach has achieved an acceptable F-measure, recall, and precision. The literature review has shown that BBNs have been deployed successfully in many applications, including military applications by Johansson and Falkman (2008), risk analysis applications by Calviño *et al.* (2016), and medical diagnosis applications by Bandyopadhyay *et al.* (2015). Our study has demonstrated that the BBN can also be successfully applied and exploited to support NLP applications for the Arabic language, and more specifically extraction of named entities from medical doamins.

- The application of BBN to analyse modern standard Arabic (MSA) texts

This study has explored a new dimension by applying BBN to process MSA texts. It is hoped that this approach may contribute to alleviating some of the challenges presented by Arabic morphology and to ensure that Arabic NLP continues to develop until the amount and quality of research is on a par with similarly important languages. In the literature, BBN has been applied to MSA for various purposes,

including handwriting recognition by Jayech *et al.* (2016), to expand the Arabic WordNet semi-automatically by Rodríguez *et al.* (2008), and to identify non-referential pronouns in Arabic texts by Hammami *et al.* (2010). Our study shows that BBN can be deployed successfully to analyse MSA text in the context of NER.

- The application of BBN to the extraction of complex medical entities

The BBN approach has been applied to many diagnostic medical domains but not to the extraction of medical entities from texts. Our BBN approach is a significant contribution to processing and learning and recognising entities related to Arabic medical texts. In the literature, most of the Arabic NER systems focused on extracting NEs related to the news domain. The application of our study is to the medical domain with the aim of extracting complex entities related to the cancer domain.

- The production of a manually annotated medical corpus in MSA

As has been noted, there is a distinct scarcity of both Arabic annotated corpora and medical annotated corpora. Manual annotation is by far the most accurate method of preparing corpora for NLP research. We believe that our manually annotated corpora in Arabic of a well-known medical corpus is an important and valuable contribution to the body of Arabic NER research.

- The evaluation of AMIRA tool (tokeniser and POS tagger)

This study has also conducted and published an evaluation of the performance of AMIRA tools, which is a valuable contribution to the researchers involved in processing Arabic texts. Although the AMIRA toolkit is essential for Arabic NER, it had not yet been tested against Arabic digital medical documents.

- Measuring the impact of using different features alongside BBN on the NER task.

As research progresses into new languages and domains, the number of features and toolkits accessible becomes much broader. However, not all these features are transferable between domains and languages. Each language and domain has its own specific set of effective features. It is necessary to evaluate the effectiveness of these features. This study has shed light on how the features selection process is important and on its impact on the recognition task.

- Assessing the impact of using different sliding window sizes on the performance of the NER task.

Window sizes are important to understanding how different Named Entities can be detected. Due to the agglutinative nature of Arabic language, certain Named Entities may be surrounded by other

words that can be used to detect them. Furthermore, medical Named Entities are often composed of multiple words, which requires a broader window for detection. As evidenced by the lower success rate for detecting Treatment Named Entities as opposed to Disease Named Entities, this research has demonstrated that different window sizes are needed for different Named Entity types.

6.4 Conclusion and Future Directions

In any domain in which it is used, NER has numerous applications, and medical texts are no exception. Applied to the medical domain, NER can assist in the detection of patterns in medical records, allowing doctors to make better diagnoses and treatment decisions, enabling different medical staff to quickly assess a patient's records and ensure that patients are informed about their data, as just a few examples. However, all these applications would require a very high level of accuracy. To improve the accuracy of NER in this domain, new approaches need to be developed that are tailored to the types of NEs to be extracted and categorised.

In an effort to solve this problem, our research applied the BBN to the process. A probabilistic model for prediction of random variables and their dependencies, BBNs are used in other fields to detect and explain patterns. Our research is the first to apply the BBN to NER. The BBN sets itself apart from other machine-learning algorithms in that it can establish relationships between nodes. This also means that BBNs are highly flexible. Our BBN strategy has achieved good accuracy for NEs in the classes of disease and treatment method. However, the average word length of the other two observed NE classes, diagnosis method and treatment, may have had a negative effect on their accuracy. This reduction in accuracy for named entities expressed in a number of tokens necessitates a different approach. Entities describing disease and treatment method are often composed of one to three words, whereas symptoms and treatments are expressed usually in terms of a full sentence. One way to improve the recognition of these longer entities is to specify a minimum threshold of the number of tokens associated with a given entity, and to adjust this threshold depending on the named entity class or category. Another way is to analyse these long entities at the pre-processing stage and feature extraction step and develop an algorithm to either break them into single components or replace them by a higher abstraction level.

Overall, the application of the BBN to Arabic medical NER is successful, but more development is needed to improve the accuracy to a standard at which the results can be applied to real medical systems. These future developments can be summarised as follows:

- Improving the performance of the system by adding additional significant features. For instance, the actual words themselves in the corpus can be used as a feature. This would help boost the system performance, as many NEs are expressed using certain words.

- Improving the performance of the system by building large domain-specific gazetteers. Building gazetteers can be time-consuming, However, they can significantly improve the performance of any given NER system.
- Improving the system performance by applying other Bayesian networks, such as tree augmented naive Bayes.
- Improving the BBN results where longer NEs are concerned by incorporating token thresholds proportionate to each category's typical token count or including further pre-processing steps.

The domain scope of our experiment was related to cancer. Hence, another future route to consider is to generalise our work to the medical domain. This can be achieved by expanding and improving our corpus by increasing its size to include a significant number of texts related to other diseases, annotating and recognising other entities like drugs, side-effects, and risk factors, and mining the new corpus to extract more relevant features.

In conclusion, our study has accomplished many achievements which are as follows. We have fulfilled our objectives, firstly surveying the currently available Arabic NER systems and methodologies and identified several flaws which we attempted to address in our research. We have developed an effective BBN approach to named entity recognition for the Arabic cancer domain, validated later via a k-cross fold validation approach. We have built and annotated a new corpus consisting of Arabic medical text extracted from the King Abdullah Bin Abdulaziz Arabic Health Encyclopaedia (KAAHE) website and tested our new approach on this corpus. This strengthen the claim of our novel BBN application in the context of Arabic medical NER. Notable also was our application of BBN to the analysis of modern standard Arabic, and our application of BBN to extracting complex medical entries as opposed to the traditional extraction of single entities some of which aided by capitalisation or numeric digits or acronyms. This is a small but important step towards the expansion of Arabic NER.

References list

- Abdallah, S., Shaalan, K., & Shoaib, M. (2012) Integrating rule-based system with classification for Arabic named entity recognition. In: A Gelbukh, ed. *2012 Computational Linguistics and Intelligent Text Processing*, Berlin Heidelberg, (7181), pp.11–322.
- AbdelRahman, S., Elarnaoty, M., Marwa M., & Fahmy, A. (2010) Integrated machine learning techniques for Arabic named entity recognition. *International Journal of Computer Science Issues (IJCSI)*, pp.27–368.
- Abdul-Hamid, A. & Darwish, K. (2010) Simplified feature set for Arabic named entity recognition. In: *Proceedings of the 2010 Named Entities Workshop*, Stroudsburg, PA, pp.110–115.
- Aboaoga M., & Aziz M. (2013) Arabic person names recognition by using a rule based approach. *Journal of Computer Science*, (9), pp.922–927.
- Algahtani, S. (2011) *Arabic Named Entity Recognition: A Corpus-Based Study. Ph.D. thesis*, The University of Manchester, UK.
- Al-Jumaily, H., Paloma, M., Jos, M., & Erik G. (2012) A real time named entity recognition system for Arabic text mining. *Language Resources and Evaluation Journal*, pp.543–563.
- Alkharashi, I. (2009) Person named entity generation and recognition for Arabic language. In: *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, pp.205–208.
- Al-Shalabi, R., Ghassan K., Al-Sarayreh, B., Khanfar, K., AlGhonmein, A. Talhouni, H., & 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.
- Al-Shoukry, S., & Omar, N. (2015). Proper nouns recognition in Arabic crime text using machine learning approach. *Journal of Theoretical and Applied Information Technology*, 79(3), 506.
- Alhawarat, M. (2015). Using N—Grams and Simple Rules. *Asian Journal of Information Technology*, 14(8-12), 287-293.
- Alotaibi, F., & Lee, M. G. (2013). Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by utilizing Wikipedia. In *IJCNLP*(pp. 392-400).
- Alotaibi, F., & Lee, M. G. (2014). A Hybrid Approach to Features Representation for Fine-grained Arabic Named Entity Recognition. In *COLING* (pp. 984-995).

- Alotaibi, F. (2015). *Fine-grained Arabic named entity recognition* (Doctoral dissertation, University of Birmingham).
- Alruily, M. (2012). *Using text mining to identify crime patterns from Arabic crime news report corpus*. (Doctoral thesis, De Montfort University)
- Alsughayr A. (2013) 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*; 1: 53-4
- Althobaiti, M. (2016). *Minimally-supervised methods for Arabic Named Entity Recognition* (Doctoral dissertation, University of Essex).
- Althobaiti, M., Kruschwitz, U., & Poesio, M. (2013) *A Semi-supervised Learning Approach to Arabic Named Entity Recognition*. University of Essex, UK.
- Amor, N. B., Benferhat, S., & Elouedi, Z. (2004,). Naive bayes vs decision trees in intrusion detection systems. In *Proceedings of the 2004 ACM symposium on Applied computing* (pp. 420-424). ACM.
- Anderson, Stephen R. (2003). Morphology. *Encyclopedia of Cognitive Science*, MacMillan. vol. III, pp. 78-83; "Leonard Bloomfield," vol. I, pp. 402-405.
- Ang, S., Ong, H., & Low, H. (2016). Classification Using the General Bayesian Network. *Pertanika Journal of Science & Technology*, 24(1).
- Aulakh, N., and Kaur, Y. (2014) Review Paper on Name Entity Recognition of Machine Translation. *International Journal of Advanced Research in Computer Science and Software Engineering* Volume 4(4), pp. 503-508
- Asharef, M., Omar, N., & Albared, M. (2012) Arabic Named Entity Recognition in Crime Documents. *Journal of Theoretical & Applied Information Technology*, (44), pp. 1-6.
- Attia, M. A. (2008). *Handling Arabic morphological and syntactic ambiguity within the LFG framework with a view to machine translation* (Doctoral dissertation, University of Manchester).
- Babych, B., & Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT* (pp. 1-8). Association for Computational Linguistics.
- Bach, E. (1986). Natural language metaphysics. *Studies in Logic and the Foundations of Mathematics*, 114, 573-595.

- Baker, P. (2006). *Using corpora in discourse analysis*. A&C Black.
- Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburgh University Press.
- Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrissi, M., & O'Connor, P. (2015). Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29(4), 1033-1069.
- Baron, A., Rayson, P., & Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1), 41-67.
- Benajiba, Y., Zitouni, I., Diab, M., & Rosso, P. (2010). Arabic named entity recognition: using features extracted from noisy data. In *Proceedings of the ACL 2010 conference short papers* (pp. 281-285). Association for Computational Linguistics
- Benajiba, Y., Diab, D., & Paolo R. (2009) Arabic named entity recognition: A feature-driven study. *IEEE Transactions on Audio, Speech, and Language Processing*, pp.926–934.
- Benajiba, Y., & Paolo, R. (2008) 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.
- Benajiba, Y., Diab, M., & Paolo R. (2008) Arabic named entity recognition: An SVM-based approach. In: *Proceedings of Arab International Conference on Information Technology (ACIT)*, Hammamet, pp.16–18.
- Benajiba, Y., & Paolo, R. (2007) 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)*, Mumbai, pp.1814–1823.
- Benajiba, Y., Rosso, P., & Benedíruiz, J. M. (2007). Anersys: An arabic named entity recognition system based on maximum entropy. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 143-153). Springer Berlin Heidelberg.
- Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. University of Michigan Press.
- Benzenberg (2014) Concordance Software in Second Language Instruction. Retrieved from https://www.academia.edu/6846244/Concordance_Software_in_Second_Language_Instruction

- Bidhendi, M., Behrouz M., & Hosein Jouzi. (2012) 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.
- Bilmes, J. A. (2004) Graphical models and automatic speech recognition. In: *Mathematical foundations of speech and language processing*, Springer New York, pp. 91-245.
- Blackburn, P., & Bos, J. (2003). Computational semantics. *Theoria: An International Journal for Theory, History and Foundations of Science*, 27-45.
- Bodenreider, O., & Zweigenbaum, P. (2000). Identifying proper names in parallel medical terminologies. *Studies in health technology and informatics*, 77, 443.
- Bodnari, A., Deleger, L., Laverigne, T., Neveol, A., & Zweigenbaum, P. (2013). A Supervised Named-Entity Extraction System for Medical Text. In *CLEF (Working Notes)*.
- Brannen, J. (2005) NCRM Methods Review Papers, NCRM/005. *Mixed Methods Research: A discussion paper*. URL: < <http://bcs.org/upload/pdf/cop.pdf>> (Accessed April 2014)
- Bryman, A. (2012) *Social research methods*. Oxford university press. ISO 690.
- Buckwalter T. (2002) *Buckwalter Arabic Morphological Analyzer Version 1.0* Linguistic Data Consortium, University of Pennsylvania.
- Calviño, A., Grande, Z., Sánchez-Cambronero, S., Gallego, I., Rivas, A., & Menéndez, J. (2016). A Markovian–Bayesian Network for Risk Analysis of High Speed and Conventional Railway Lines Integrating Human Errors. *Computer-Aided Civil and Infrastructure Engineering*, 31(3), 193-218.
- Chinchor, N., & Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding*(p. 29).
- Chinchor, N., Robinson, P., & Brown, E. (1998). Hub-4 Named Entity task definition version 4.8. Available by ftp from www.nist.gov/speech/hub4_98.
- Chowdhury, G. (2003) Natural language processing. *Annual review of information science and technology*, 37(1), 51-89.
- Cooper, G., (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2-3), 393-405.

- Crossley, S., Salsbury, T., McNamara, D., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly*, 45(1), 182-193.
- Crystal, D. (2004). *The Cambridge encyclopedia of the English language*. Ernst Klett Sprachen.
- Daly, R., Shen, Q., & Aitken, S. (2011). Learning Bayesian networks: approaches and issues. *The knowledge engineering review*, 26(02), 99-157.
- Darwish, K., & Gao, W. (2014). Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In *LREC* (pp. 2513-2517).
- Diab, M. (2009) Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging and Base Phrase Chunking. *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, 2009.
- Diab, M, Hacioglu, K., & Jurafsky, D. (2007) Arabic Computational Morphology: Knowledge-based and Empirical Methods, chapter Automated Methods for Processing Arabic Text: *From Tokenization to Base Phrase Chunking*. Kluwer/springer edition
- Dijk, T. A. (1983). Discourse analysis: Its development and application to the structure of news. *Journal of communication*, 33(2), 20-43.
- Easterby-Smith, M., Thorpe, R., & Jackson, P. R. (2012). *Management research*. Sage.
- Elsebai, A. (2009) *A Rules Based System for Named Entity Recognition in Modern Standard Arabic*, *PhD thesis*. University of Salford, UK.
- Elsebai, A., & Meziane, F. (2011). Extracting person names from Arabic newspapers. In *Innovations in Information Technology (IIT), 2011 International Conference on* (pp. 87-89). IEEE.
- Elsebai, A., Meziane, F., & Belkredim, F. Z. (2009). A rule based persons names Arabic extraction system. *Communications of the IBIMA*, 11(6), 53-59.
- Etzioni, O., Cafarella, M., Downey, D., Popescu, A. M., Shaked, T., Soderland, S., & Yates, A. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1), 91-134.
- Farber, B., Dayne F., Habash, H. & Owen, R. (2008) Improving NER in Arabic using a morphological tagger in *proceedings of the Sixth International Conference on Language Resources and Evaluation*. (LREC2008), pages 2,509–2,514, Marrakech.

- Farghaly, A., & Shaalan, K. (2009) Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing (TALIP)*, pp.1–22.
- Feldman, S. (1999) NLP meets the jabberwocky. Online, 23, 62-72.
- Firth, J. (1957) *Papers in Linguistics, 1934-1951*. Oxford: Oxford University Press.
- Friedman, N., & Goldszmidt, M. (1996). Building classifiers using Bayesian networks. In *Proceedings of the national conference on artificial intelligence* (pp. 1277-1284).
- Fry, J. (2011), Tokenizing, lecture notes distributed in Linguistics 497: Corpus Linguistics, Spring 2011, Boise State University
- Habash N. (2010) *Introduction to Arabic Natural Language Processing*. Synthesis Lecture on Human Language Technologies. A Publication in the Morgan & Claypool Publishers series, UAS.
- Habash, N., Rambow, O., & Roth, R. (2009) MADA+TOKAN: A toolkit for Arabic tokenization diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR)*, Cairo, Egypt
- Hale, R. (2005) Text mining: getting more value from literature resources. *Drug Discov. Today* 10, 377–379
- Hammami, S., Sallemi, R., & Belguith, L. (2010). A bayesian classifier for the identification of non-referential pronouns in arabic. In *Informatics and Systems (INFOS), 2010 The 7th International Conference on* (pp. 1-6). IEEE.
- Harvey, K. (2013). *Investigating Adolescent Health Communication: A Corpus Linguistics Approach*. Bloomsbury Publishing.
- Haug, P., Koehler, S., Christensen, L., Gundersen, M. & Van Bree, R. (2001) *Probabilistic method for natural language processing and for encoding free-text data into a medical database by utilizing a Bayesian network to perform spell checking of words*, U.S. Patent No. 6,292,771. Washington, DC: U.S. Patent and Trademark Office.
- Heckerman, D. (1998) A tutorial on learning with Bayesian networks. Springer Netherlands.
- Hoey, M. (1991) *patterns of lexis in text*. Oxford university press.

- Ide, N., & Véronis, J. (1998). Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1), 2-40.
- Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI* (Vol. 7, pp. 1606-1611).
- Gledhill, C., (2000). *Collocations in Science Writing*. Tübingen, Gunter Narr, 7-20
- Goweder, A., Poesio, M., De Roeck, A. N., & Reynolds, J. (2004). Identifying Broken Plurals in Unvowelised Arabic Tex. In *EMNLP* (pp. 246-253).
- Guo, J., Gu X., Xueqi C., & Hang Li. (2009) 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.
- Jammalamadaka, A. K. (2004). Aspects of inference for the Influence Model and related graphical models (Doctoral dissertation, Massachusetts Institute of Technology).
- Jayech, K., Mahjoub, M., & Essoukri Ben Amara, N. (2016). Arabic handwritten word recognition based on dynamic bayesian network. *Int. Arab J. Inform. Technol.(IAJIT)*, 13(3).
- Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC bioinformatics*, 9(Suppl 3), S3.
- Jochim, C., Sacaleanu, B., & Deleris, L. (2014). Risk Event and Probability Extraction for Modeling Medical Risks. In *2014 AAAI Fall Symposium Series*.
- Johansson, F. & Falkman, G. (2008) A Bayesian network approach to threat evaluation with application to an air defense scenario. In: *11th International Conference on Information fusion*, Cologne, Germany.
- Jurafsky, D., (2003) Pragmatics and Computational Linguistics. In Laurence R. Horn & Gregory Ward (eds.) *Handbook of Pragmatics*. Oxford: Blackwell.
- Karov, Y., & Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational linguistics*, 24(1), 41-59.
- Kazama, J., & Torisawa, K. (2008). Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. *Proceedings of ACL-08: HLT*, 407-415.

- Khoja, S. (1999) *Stemming Arabic Text*. Computing Department, Lancaster University, Lancaster, U.K
- Khoja, S., Garside R., & Knowles, G. (2001). A tag set for the morphosyntactic tagging of Arabic. In *Proceedings of Corpus Linguistics*, Lancaster, UK.
- Kilgariff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, 105, 116
- Kothari, C. (2004) *Research Methodology; Methods and Technique Dharmesh Printers*. New Delhi, India
- Koulali, R., & Abdelouafi, M. (2012) A contribution to Arabic named entity recognition. In: *Proceedings of 10th International Conference on ICT and Knowledge Engineering*, Morocco, pp.46–52.
- Krallinger, M., & Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. *Genome biology*, 6(7), 1.
- Leaman, R., & Gonzalez, G. (2008). BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific symposium on biocomputing* (Vol. 13, pp. 652-663).
- Lee, C., Hwang, Y. G., Oh, H. J., Lim, S., Heo, J., Lee, C. H., ... & Jang, M. G. (2006, October). Fine-grained named entity recognition using conditional random fields for question answering. In *Asia Information Retrieval Symposium* (pp. 581-587). Springer Berlin Heidelberg.
- Lee, H., Hsu, Y., & Kao, H. (2015). An enhanced CRF-based system for disease name entity recognition and normalization on BioCreative V DNER Task. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* (pp. 226-233).
- Lee, S., & Abbott, P. A. (2003). Bayesian networks for knowledge discovery in large datasets: basics for nurse researchers. *Journal of biomedical informatics*, 36(4), 389-399.
- Leung, K. M. (2007). Naive Bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*
- Liddy, E. (1998). Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science*, 24, 14-16.
- Liddy, E. (2001) Natural Language Processing. In: *Encyclopaedia of Library and Information Science*, 2nd ed. Marcel Decker, Inc., New York.

- Lin, Y., & Druzdel, M. J. (1997). Computational advantages of relevance reasoning in Bayesian belief networks. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence* (pp. 342-350). Morgan Kaufmann Publishers Inc.
- Liu, J., Shang, J., Wang, C., Ren, X., & Han, J. (2015). Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 1729-1744). ACM
- Maloney, J., & Niv, M. (1998). 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.
- Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing* (Vol. 999). Cambridge: MIT press.
- Marsh, E., & Perzanowski, D. (1998). MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the seventh message understanding conference (MUC-7)* (Vol. 20).
- Menner, T., Höpken, W., Fuchs, M., & Lexhagen, M. (2016). Topic Detection: Identifying Relevant Topics in Tourism Reviews. In *Information and Communication Technologies in Tourism 2016* (pp. 411-423). Springer International Publishing.
- Meselhi, M. A., Bakr, H. M. A., Ziedan, I., & Shaalan, K. (2014). A Novel Hybrid Approach to Arabic Named Entity Recognition. In *China Workshop on Machine Translation* (pp. 93-103). Springer Berlin Heidelberg.
- Mohammed. N, & Omar, N. (2012) Arabic Named Entity Recognition Using Artificial Neural Network. *Journal of Computer Science*, pp.1285-1293.
- Mohit, B., Schneider, N., Bhowmick, R., Oflazer, K., & Smith, N. A. (2012,). Recall-oriented learning of named entities in Arabic Wikipedia. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 162-173). Association for Computational Linguistics.
- Morsi, A., & Rafea, A. (2013). Studying the impact of various features on the performance of Conditional Random Field-based Arabic Named Entity Recognition. In *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on* (pp. 1-5). IEEE.
- Nadeau, D. & Satoshi S. (2007) A survey of named entity recognition and classification. *Lingvisticae Investigationes*, pp.3–26.

- Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov, V. (2016). SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016)*, San Diego, US (forthcoming).
- Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K., & Ay-shanker, K. V. (2003). A biological named entity recognizer. In *Pac Symp Biocomput* (p. 427).
- Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, 24(2), 223-242.
- Nezda, L, Andrew H, John L, & Sarmad Fayyaz. (2006) 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-06)*, Genoa, pp.41–46.
- Niedermayer, D. (2008). An introduction to Bayesian networks and their contemporary applications. In *Innovations in Bayesian Networks* (pp. 117-130). Springer Berlin Heidelberg.
- Nikovski, D. (2000) Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *Knowledge and Data Engineering, IEEE Transactions on*, 12(4), pp.509-516.
- Orita, N. (2015). *Computational modeling of the role of discourse information in language production and acquisition* (Doctoral dissertation, University of Maryland, College Park).
- O'Steen, D., & Breeden, D. (2009) *Named Entity Recognition in Arabic: A Combined Approach*. BA (Hons), Stanford University.
- Oudah, M., & Shaalan, K. (2012). A Pipeline Arabic Named Entity Recognition using a Hybrid Approach. In *COLING* (pp. 2159-2176).
- Oudah, M., & Shaalan, K. (2013) Person name recognition using the hybrid approach. *Natural Language Processing and Information Systems*, Berlin Heidelberg, (7934), pp.237–248.
- Pasha, A., Al-Badrashiny, M., Diab, M. T., El Kholy, A., Eskander, R., Habash, N., ... & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *LREC* (Vol. 14, pp. 1094-1101).
- Peng, F., Feng, F., & McCallum, A. (2004). Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics* (p. 562). Association for Computational Linguistics.

- Pearl, J. (1998). *Probabilistic reasoning in intelligent systems: Networks for plausible inference*. San Francisco: Morgan Kaufman.
- Popowich, F. (2005). Using text mining and natural language processing for health care claims processing. *ACM SIGKDD Explorations Newsletter*, 7(1), 59-66.
- Rindflesch, T. C., Tanabe, L., Weinstein, J. N., & Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*(p. 517). NIH Public Access.
- Roberts, A.; Al-Sulaiti, L., Atwell, E. (2005). aConCorde: towards a proper concordance of Arabic In: *Proceedings of Corpus Linguistics 2005*.
- Robson, C., (2002), *Real world research*, 2nd, Blackwell, Oxford
- Rodríguez, H., Farwell, D., Ferreres, J., Bertran, M., Alkhalifa, M., & Martí, M. A. (2008). Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. In *LREC*.
- Roth, D., & Yih, W. T. (2002). Probabilistic reasoning for entity & relation recognition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1 (pp. 1-7)*. Association for Computational Linguistics.
- Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Samy, D., Moreno-Sandoval, A., Bueno-Diaz, C., Garrote-Salazr, M., & Guirao, J. (2012) Medical Term Extraction in an Arabic Medical Corpus. *Proceedings of the 8th Language Resources and Evaluation Conference*. Istanbul, Turkey.
- Samy, D., Moreno, A., & Guirao, J. (2005). A proposal for an Arabic named entity tagger leveraging a parallel corpus. In *International Conference RANLP, Borovets, Bulgaria* (pp. 459-465).
- Saunders, M., Lewis, P., & Thornhill, A. (2009) *Research methods for business students* fifth edition. Essex, UK, Financial Times/ Prentice Hall.
- Sawalha, M. & Atwell, E. (2009) Linguistically Informed and Corpus Informed Morphological Analysis of Arabic. In: *Proceedings of the 5th International Corpus Linguistics Conference CL2009*, 20-23 July 2009, Liverpool, UK.

- Schubert, L., (2015) Computational Linguistics. *The Stanford Encyclopedia of Philosophy* (Spring 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2015/entries/computational-linguistics/>.
- Seale, C., & Charteris-Black, J. (2010). Keyword analysis: A new tool for qualitative research. *The SAGE handbook of qualitative methods in health research*. London: Sage, 536-556.
- Shaalán, K. (2014) A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics*, 40:2, MIT Press.
- Shaalán, K. (2010) Rule-based Approach in Arabic Natural Language Processing. *The International Journal on Information and Communication Technologies (IJICT)*, pp.11-19.
- Shaalán, K., & Raza, H. (2008) Arabic named entity recognition from diverse text types. *Advances in Natural Language Processing*, (5221), pp.440–451.
- Shaalán, K., & Raza, H. (2007). Person name entity recognition for Arabic. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources* (pp. 17-24). Association for Computational Linguistics.
- Shabat, H., & Omar, N. (2015). Named Entity Recognition in Crime News Documents Using Classifiers Combination. *Middle-East Journal of Scientific Research*, 23(6), 1215-1221.
- Shihadeh, C., & Neumann, G. (2012) ARNE: A tool for named entity recognition from Arabic text. In: *Fourth Workshop on Computational Approaches to Arabic Script-based Languages (CAASL4)*, San Diego, CA, pp.24–31.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
- Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1), 1-38.
- Sondhi, P., (2010) A Survey on Named Entity Extraction in the Biomedical Domain. University of Illinois at Urbana Champaign
- Steedman, M. (1996). Natural language processing. In M. Boden (Ed.), *Artificial Intelligence*, (pp 229-266). San Diego, CA: Academic Press.
- Sundheim, B. (1996) Overview of results of the MUC-6 evaluation. *Proceedings of a workshop on held at Vienna*, 6-8 May, pp.423-442.

- Ticehurst, J., Newham, L., Rissik, D., Letcher, R., & Jakeman, A. (2007) Bayesian network approach for assessing the sustainability of Coastal Lakes in New South Wales, Australia. *Environmental Modelling and Software*, pp.1129-1139.
- Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4* (pp. 142-147). Association for Computational Linguistics.
- Traboulsi, H. (2009) 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.
- Tsuruoka, Y., & Tsujii, J. I. (2003). Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine-Volume 13* (pp. 41-48). Association for Computational Linguistics.
- Van Roey, J. (1990). *French-English contrastive lexicology: An introduction*(Vol. 14). Peeters Publishers.
- Velikova, M., van Scheltinga, J., Lucas, P., & Spaanderman, M. (2014) Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1), pp.59-73.
- Voutilainen, A. (2003) Part-of-speech tagging. In R. Mitkov, editor, *The Oxford handbook of computational linguistics*. University Press, Oxford, pp. 219–232.
- Wang, J., Peng, Y., Liu, B., Wu, Z., Deng, L., & Jiang, T. (2016). Extracting Clinical entities and their assertions from Chinese Electronic Medical Records Based on Machine Learning.
- Williams, C. (2007) Research Methods. *Journal of Business & Economic Research*, (5).
- Wooldridge, S. (2003) *Bayesian Belief Networks*. Centre for Complex System Science, CSIRO, Canberra.
- Wu, Y., Jiang, M., Lei, J., & Xu, H. (2015). Named entity recognition in Chinese clinical text using deep neural network. *Studies in health technology and informatics*, 216, 624.
- Yang, Y. (1994, August). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In: *Proceedings of the 17th annual international ACM SIGIR*

- conference on Research and development in information retrieval*, Springer-Verlag New York, Inc., pp. 13-22.
- Zaghouani, W., Pouliquen, B., Ebrahim, M., & Steinberger, R. (2010) Adapting a resource-light highly multilingual named entity recognition system to Arabic. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, Valletta, pp.563–567.
- Zaghouani, W. (2012) RENAR: A rule-based Arabic named entity recognition system. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(1):2:1–2:13.
- Zayed, O., & El-Beltagy, S. (2012) Person name extraction from modern standard Arabic or colloquial text. In: *Proceedings of the 8th International Conference on Informatics and Systems conference*, Cairo, pp.44–48.
- Zhang, L.: Text2Ngram. <http://homepages.inf.ed.ac.uk/s0450736/ngram.html>
- Zhang, S., & Elhadad, N. (2013). Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6), 1088-1098.
- Zirikly, A., & Diab, M. (2015). Named entity recognition for arabic social media. In *Proceedings of naacl-hlt* (pp. 176-185).
- Zou, M. & Conzen, S. (2005) A New Dynamic Bayesian Network (DBN) Approach for Identifying Gene Regulatory Networks from Time Course Microarray Data. *Bioinformatics*, 21(1), pp.71-79.

Appendix I. Stopwords list

إِذَا	ف	إِنْ	أَنْ	مَا	كَانَ
اليوم	أُخْرَى	شَانِع	هَذِهِ	أَنْهَا	أَخْرَ
بَعْضُ	أَجَلَ	هِيَ	بَعْدَ	كَانَتْ	الَّذِي
هَذَا	قَدْ	أَيَّةُ	يُمْكِنُ	أَوْ	إِلَى
أَيُّ	مِنْ	الْمُخْتَلَفَةِ	يَكُونُ	أَكْثَرُ	يُسَاعِدُ
عَلَى	أَفْضَلَ	شَكْلَ	فِرَاشَةٍ	العديد	التي
تسمى	ك	مَا	يُمْكِنُ	أَنْ	أَيْضًا
حَيْثُ	فِي	الْغَالِبِ	لَا	أَسْبُوعَيْنِ	وَاحِدَةٍ
مِنْ	رَئِيسِي	ذَلِكَ	عِنْدَ	الَّذِينَ	أَمَّا
يَلِي	بِيدَ	سَوَى	غَيْرِ	لَا سِيَمَا	مَتَى
أَنْيَ	أَيُّ	أَيَّانَ	أَيْنَ	بِكُمْ	بِمَا
بِمَاذَا	بِمَنْ	كَمْ	كَيْفَ	مَا	مَاذَا
مَتَى	مِمَّا	مِمَّنْ	مِنْ	حَيْثُمَا	كَيْفَمَا
مَا	مَتَى	مِنْ	مَهْمَا	أَوَّلُنْكَ	أَوَّلُنْكُمْ
تلك	تلكم	تلكما	ثُمَّ	ثُمَّةُ	ذَا
ذَاكَ	ذَلِكَ	ذَلِكُمْ	ذَلِكُمَا	ذِي	كَذَلِكَ
هؤلاء	هَاهُنَا	هَذَا	هَذَانِ	بَعْضُ	هَكَذَا
هنا	هَنَّاكَ	هَنَّاكَ	أَيُّ	إِذْ	لَكِنَّهُ
جدا	إِذَا	بَعْضُ	تَجَاهُ	تَلْقَاءُ	جَمِيعُ
حسب	حَيْثُ	سَبْحَانَ	سَوَى	شَبْهَ	كُلِّ
لِعمر	لَمَّا	مِثْلُ	مَعَاذَ	مَعَ	نَحْوُ
التي	أَمَامَكَ	الَّذِي	أَمِينِ	أَكْثَرُ	أَقْلُ
اللاتي	اللاتي	اللتان	اللتيا	اللتين	اللذان
اللذين	اللواتي	ذَا	ذَاتَ	مَا	أَبَ
أَخَ	حَمَ	ذَوِ	فَوِ	لَنْ	لَوْ
لولا	لَوْمًا	نَعَمْ	إِنْ	لَاتَ	مَا
لا	أَنْ	عَلِ	كَانَ	لَعَلَّ	كَيْ
أجمع	جَمِيعُ	عَامَةً	عَيْنِ	كُلِّ	كَلَّا
كلاهما	كَلْتَا	كَلَيْكُمَا	كَلَيْهِمَا	نَفْسِ	إِلَّا
حاشا	خَلَا	عَدَا	لَكِنْ	فِيمَ	فِيمَا
هل	سَوْفَ	كَمَا	لَكِي	لَكَيْلَا	رَبِّ
على	عَنْ	فِي	مِنْذَ	مِنْذَ	لَمْ
لَمَّا	أَجَلَ	إِذَنْ	إِي	بَلَى	جَلَلَ

Appendix II. Gazetteers

Diseases Entity						
سرطان	اللويميا	ابيضاض	الايضااض	لوكيميا	السااركومة	سرطانات
سرطانة	السرطانة	ساركومة	الكارسينومة	الحرشقية	القاعدية	السحانية
الأرومية	الأرومي	القوائم	المبيضي	المريني		
Symptoms Entity						
الحمى	القشعريرة	الوهن	التعب	نقص	ضخامة	النزف
الكدمات	التعرق	الصداع	القيء	اضطراب	الاختلاج	تقرحات
كتلة	الضعف	الألم	ألم	تقيؤ	غثيانا	تخلخل
نزف	تورم	الصداع	خدر	تنميل	اصفرار	الإنهاك
وجع	انزعاج	صعوبة	مشاكل	تشنجات	السعال	سعال
التهاب	تشوش	احمرار	الحكة	انتفاخ	شحوب	
Treatment Methods Entity						
المعالجة	الجراحة	الكيميائية	زرع	السريية	الكماوي	الشعاعية
الإشعاعي	الاستئصال	التخثير	الدعامة	التبريد	الليزر	المستهدفة
Diagnosis Methods Entity						
فحصا	المجهر	خزعة	الرنين	المقطعي	منظار	المنظار
الخزعات	التصويرية	تنظير	البوزيتروني	لطاخة	مسابر	مسابر
الصوتية	السينية	المغناطيسي	الاختزاع			

Appendix III. Lexical Markers

Diseases Entity						
سرطان	السرطانة	ابيضاض	الايبيضاض	ساركومة	الساركومة	سرطانات
سرطانة	الأرومي					
Symptoms Entity						
الحمى	ظهور	وجود	كتلة	نقص	ضخامة	النزف
الكدمات	التعرق	تشوش	التهاب	اضطراب	احمرار	تقرحات
مفرزات	شحوب	الألم	ألما	تقيؤا	غثيانا	تخلخل
نزف	تورم	صعوبة	خدر	تنميل	اصفرار	سعال
وجع	انزعاج		مشاكل	تشنجات	انتفاخ	حكة
Treatment Methods Entity						
المعالجة	التجارب	تجارب	زرع	فغر	التبريد	التخثير
الإشعاعي	الاستئصال					
Diagnosis Methods Entity						
فحوصا	الصورة	خزعة	الرنين	لطاخة	منظار	الأمواج
المقطعي						

