Resolving Ambiguous Entity through Context Knowledge and Fuzzy Approach

Hejab M. Alfawareh
Dept of Information Systems
Faculty of Science & Information Technology
Zarqa University, Zarqa, Jordan
hejab@zpu.edu.jo

Shaidah Jusoh
Dept. of Computer Science
Faculty of Science & Information Technology
Zarqa University, Zarqa, Jordan
shaidah@zpu.edu.jo

***Abstract*—Entity extraction is considered as a fundamental step in many text mining applications such as machine translation, text summarization and text categorization. However, the major challenging issue in extracting the entity from a sentence is the ambiguity problem, namely lexical ambiguity. While a human has a cognitive capability to resolve the meaning easily based on his/her knowledge, it is very difficult for a machine to do so. This paper proposed a new technique for resolving the ambiguity problem through a fuzzy approach and context knowledge. The technique integrates subject and lexical knowledge, the possibility theory, and fuzzy sets into natural language processing. Lexical knowledge was obtained from WordNet, while subject and lexical knowledge have been deployed as context knowledge. Possibility theory and fuzzy sets were applied to select the most possible meaning of an ambiguous entity based on the context. The work was conducted on the noun part-of-speech only. The technique was implemented and tested with 1110 sentences. Precision and recall measurement metrics were used as an evaluation metric. The obtained precision rate is 85.7% and 80.3% for recall. The results indicate that the proposed technique is successful**. *(Abstract)*

Keywords-natural language processing; ambiguity; context knowledge, fuzzy approach; information extraction

I. INTRODUCTION

Valuable information is normally embedded inside unstructured texts. Extracting the valuable information requires reading. However, this task is very time consuming. Having an automated system that can extract entities from the texts, store them in a database, and then use machine learning algorithms to manipulate those entities and produce a piece of information is desirable. The desired system requires the study of natural language processing (NLP) and information extraction (IE). NLP is a field of computer science and linguistics concerned with the interaction between computers and natural languages. Natural language understanding is sometimes referred to as an AI-complete problem because natural language understanding requires extensive knowledge about the language and the ability to manipulate it [38]. The most challenging issue in understanding is, any natural language is not free from the ambiguity problem.
On the other hand, IE is an effective way to populate the contents of a relational database. Its process turns the unstructured information embedded in texts into structured data. IE is a domain specific task; the important types of objects and events for one domain can be quite different from those in another domain [18]. Fundamental tasks in IE are named entity recognition and named entity relation extraction.

Technically, an entity recognition task focus on identifying relevant concept/entity where the criteria for relevance are predefined by a user in a form of a template that is to be filled. Only a fraction of the text contains relevant information, and a relatively simple, predefined, rigid target representation that the information is mapped into. Work template slots and their associated filling criteria has been anticipated and encoded by the system builder. However as argued by [18], IE should do more than that to make it more successful. Word-by-word match (as exist in most of the current IE systems) is not enough.

To extract valuable knowledge from texts requires a system to extract most relevant information from texts by extracting entities and facts which are distributed in texts. This activity requires an intelligent IE tool. To the best of our knowledge no robust IE tool has been developed so far. The main reason why until now we have not got the 'dream tool' is because natural language is ambiguous [10], [55]. Most of the current work in IE such as [17], [12], [15], [49], [14], [13], [9] do not really focus on the noun part-of-speech entity extraction. Nevertheless, most concepts and facts in texts are represented in the noun form. Consequently, it can be argued that extracting and classifying entities based on their meanings are necessary for text mining applications. Technically, an entity recognition task focus on identifying relevant concept/entity where the criteria for relevance are predefined by a user in a form of a template that is to be filled. Only a fraction of the text contains relevant information, and a relatively simple, predefined, rigid target representation that the information is

mapped into. Work template slots and their associated filling criteria has been anticipated and encoded by the system builder. However as argued by [18], IE should do more than that to make it more successful. A Word-by-word match (as exist in most of the current IE systems) is not enough.

The ambiguity problem in a natural language can be classified into 4 types; lexical ambiguity, structural ambiguity, semantic ambiguity, and pragmatic ambiguity [26]. Not all ambiguities can be easily identified and some of them requires a deep linguistic analysis. Ambiguity in entity extraction is a problem of lexical ambiguity. Lexical ambiguity occurs when a word has more than one meaning [35]. As stated by Zadeh (1978) "it is difficult to find a word that has only one meaning".

This research paper introduces a new technique to handle the ambiguity problem in entity extraction. Entity extraction is a process of extracting tangible objects which are normally represented by noun, pronoun, and proper noun part-of-speech. In this paper, only noun part-of-speech is presented. The proposed new technique is obtained by using context knowledge and a fuzzy approach. The new technique is integrated into natural language processing. This paper is organized as follows; Section II presents technical background and related work of the proposed technique. The proposed technique is presented in Section III. Experiments and results are presented in Section IV and V. The discussion is presented is Section VI and conclusions are made in Section VII

II. RESEARCH BACKGROUND

The most relevant research areas of the proposed technique are NLP, possibility theory and fuzzy sets, IE, and word sense disambiguation. An overview of each area and its previous work is presented in the following subsections.

A. Natural Language Processing (NLP)

NLP work focus on analyzing of human language so that computers can understand natural languages as a human being does. The ultimate goal of NLP is to develop a software program that enable computers to understand and a generate language used by humans. This field is moving rapidly and much work has been conducted in the last 10 years. Although the goal of NLP's work remain far from being success, a significant positive outcome has been shown in some research work [48], [11], [43], [24], [16]. NLP is a technology that concerns with natural language generation (NLG) [42] and natural language understanding (NLU) [45], [23]. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic [4]. NLU is independent from speech recognition [28]. However, the combination of the two may produce a powerful human-computer interaction system. When combined with NLU, speech recognition transcribes an acoustic signal into a text. Then the text is interpreted by an understanding component to extract the meaning. NLU has been an active area of research for a few decades. In NLU there are two important components: syntactic and semantic analysis [25].

Syntactic analysis is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used. Syntactic processing requires natural language grammar rules, lexicon and parsing technique. Without any one of them, a syntactic structure of the sentence cannot be obtained. *Semantic analysis* is a process of translating a syntactic structure of a sentence into a *semantic representation* that is precise and unambiguous representation of the meaning expressed by the sentence. A semantic representation allows a system to perform an appropriate task in its application domain. The semantic representation is in a formally specified language. The language has expressions for real world objects, events, concepts, their properties and relationships, and so on. Semantic interpretation can be conducted in two steps: *context independent interpretation* and *context interpretation*. Context independent interpretation concerns what words mean and how these meanings combine in sentences to form sentence meanings. Context interpretation concerns how the context affects the interpretation of the sentence. The context of the sentence includes the situation, in which the sentence is used, the immediately preceding sentences, and so on.

B. Possibility Theory and Fuzzy Sets

Possibility theory was introduced by Zadeh in 1978, in the connection with fuzzy set theory, to allow a reasoning to be carried out on imprecise or vague knowledge, making it possible to deal with uncertainties on this knowledge. The theory deals with possibility distributions of variables that are restricted by fuzzy sets [29]. Possibility theory is an important component in fuzzy set theory. It can be used to estimate the possibility for an event to occur under a certain *condition*. The condition can be translated into a *context*. In possibility theory, an event can be represented as an expression containing variables, a condition can be represented by using a

restricting fuzzy set, and at the possibility for the event to occur under the condition can be estimated by a possibility distribution function. Possibility theory can be formulated not only in terms of nested bodies of evidence, but also in terms of fuzzy sets [54].

Possibility theory is a theory that uses human common sense to estimate the possibility for an *event* to occur under a certain *condition*. In possibility theory, an event can be represented as an expression containing *variables*, a condition can be represented by using *a restricting fuzzy set*, and at the possibility for the event to occur under the condition can be estimated by a *possibility distribution function.*

C. Information Extraction (IE)

IE is an enabling technology which allows an intelligent system for retrieving valuable information and knowledge from free text to be developed. Basically, IE is a process of extracting useful information from the text and storing the information in a structured database. Then a machine learning approach can be applied to the structured data for discovering new knowledge. IE task is defined by its input and its extraction target. The input can be unstructured documents like free text that are written in natural language or the semi-structured documents that are pervasive on the Web, such as tables or itemized and enumerated lists. Programs that perform the task of IE are referred to as extractors or wrappers [9]. The first step in most IE tasks is to detect and classify all the proper names mentioned in a text; a task generally referred to as named entity recognition (NER). Reference [23] defined entity as anything that can be referred to with a proper name. A process of NER refers to the combined task of finding spans of text that constitute proper names and then classifying the entities referred to according to their type. The IE tasks aim at finding specific data in natural language texts. With IE approach, events, facts and entities are extracted before the knowledge mining process is conducted. Consequently IE allows for mining the actual information presented in the texts, rather than the limited set of tags associated to the documents [27], [36]. Unlike information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post-processing, which is crucial to many text mining applications.

According to [17], IE does a more limited task than full text understanding. He pointed that in full text understanding, all the information in the text is presented, whereas in IE, the semantic range of the output, the relations will be presented are delimited. IE systems can be developed without employing NLP techniques such as deep parsing. However, recent research in IE argued that more text understanding is required. For example, [50], [18], and [19] argued that IE should be based on understanding of the structure and meaning of the natural language in which documents are written, and the goal of IE is to accumulate semantic information from text. Consequently, extracting information from texts requires lexical knowledge, grammars describing the specific syntax of the texts to be analyzed, as well as semantics [41].

D. Word Sense Disambiguation

Word sense disambiguation WSD is a topic which is very relevant to IE and NER. WSD is a process to identify the meaning of words in a computational manner. WSD has been recognized as an AI-hard problem. A break-through in this field would have a significant impact on many relevant applications, such as Web information retrieval, improved access to Web services, IE, etc [40]. WSD has obvious relationships to other fields such as lexical semantics, whose main endeavor is to define the relationships between “word” and “meaning” and “context” [1]. WSD is also known as lexical ambiguity resolution [5], [32], [31].

Generic WSD can be divided into two groups; lexical sample and all words WSD [38]. In a lexical sample, a system is required to disambiguate a restricted set of target words usually occurring one per sentence. In this type of systems, a number of instances are labeled manually (training set) and then applied to unlabeled instances (test set). This is also known as a supervised system. In all words WSD, a system is required to disambiguate all open-class words in a text. These include nouns, verbs, adjectives and adverbs. This task requires a wide coverage of systems. Thus a supervised system can potentially suffer from the problem of data sparseness, as it is unlikely that a training set of adequate size is available for a wide coverage. This is a point where the use of external knowledge is considered for WSD. This type of systems is classified into unsupervised systems. Unsupervised systems based their disambiguation decisions on knowledge sources. According to [3] knowledge sources may belong to one of broad class: syntactic, semantic and pragmatic. Syntactic knowledge sources have to do with the role of a word within the grammatical structures of sentences. Semantic knowledge relates the word to its properties. This was demonstrated by the work of [33] where they have combined knowledge gathered from WordNet with results of an anaphora resolution algorithm. Knowledge sources include corpora (a collection of text), machine readable dictionaries, semantic networks, etc.

The use of knowledge-based approach has been emonstrated in the early WSD work. For example, references [46] and 52] used manually encoded semantic knowledge for WSD. Unfortunately, the manual creation of

knowledge resources is an expensive and time consuming effort, which must be repeated every time the disambiguation scenario changes. In recent years, existing lexical resources such as machine-readable dictionaries (MRDs) like WordNet [44], [34], [39], [37] and Oxford Dictionary of English have been applied as an external source of knowledge in WSD work. Reference [40] claimed that word senses clearly fall under the category of objects that are better described through a set of structured features. Thus they have applied structural pattern recognition approach to disambiguate word senses. In their work, graph representations of word senses are automatically generated from WordNet 2.7. Other researchers who used WordNet include [22], [2], [6].

III. PROPOSED TECHNIQUE

The proposed technique is obtained by considering the use of context knowledge and fuzzy approach in resolving ambiguity in entity extraction. The proposed technique applies lexical and subject knowledge as a context and uses fuzzy sets and possibility theory to decide the most possible meaning of the entity. The technique is integrated into natural language processing.

A. Framework

The framework of the proposed method is illustrated in Figure 1. There are steps. Each step is described as follows.

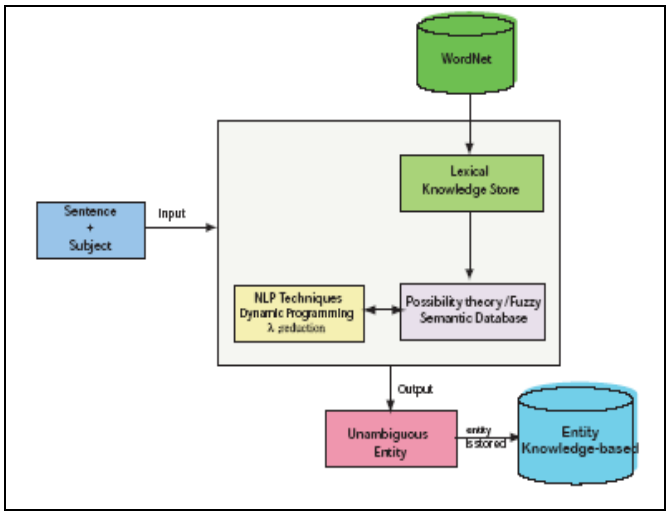

Figure 1. The framework of the proposed technique

1) **Sentence and Subject**

A word that is categorized into the noun part-of-speech is defined as an entity. A sentence consists of a sequence of entities. A sentence serves as an input. The sentence may consist of unambiguous and ambiguous entities. Although the process of extracting entities should include unambiguous and ambiguous entities, in this paper, the discussion is made for ambiguous entities only. Subject is predefined by a user of the system. The subject which is given as an input along with the sentence is a part of context knowledge. Knowledge about the context will be used in calculating the most possible semantic of an ambiguous entity.

2) **WordNet Database**

WordNet is a large lexical database of English, developed by Princeton University. The database categorized words into nouns, verbs, adjectives and adverb; each expressing a distinct concept. Nouns, verbs, adjectives and adverbs are grouped into sets of synsets. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is also freely and publicly available on the Internet for download. WordNet's structure makes it a useful tool for computational linguistics and NLP. In this work, WordNet version 3.0 is used as a reference in determining semantics of a word.

3) **Lexical Knowledge Store**

Lexical knowledge store is a knowledge-base where it contains a set of words or lexical with their semantics. The set of lexical with its semantics are extracted from the WordNet database manually. All semantics in this work are extracted from the WordNet. For instance, the word pen may have 5 possible meanings as shown in Table I. Knowledge about lexical meanings and an input subject are combined to generate context knowledge.

TABLE I. THE MEANING OF THE WORD PEN

Word (*x*)	Semantic (*sem*)
Pen	a writing tool
Pen	a livestock's enclosure
Pen	a portable enclosure for a baby
Pen	a correctional institution
Pen	a female swan

4) **Possibility theory/Fuzzy Semantic Database**

A fuzzy semantic database is introduced in this paper. It is created by utilizing the input subject, lexical knowledge, the possibility theory and fuzzy sets approach. Using the subject and lexical knowledge, the context of text is determined. The fuzzy semantic database is used in resolving ambiguous entities. Details of how fuzzy semantic database is utilized for resolving ambiguous entities will be explained in the Section III-B.

TABLE II. LEXICAL KNOWLEDDGE DATABASE IN THE KNOWLEDGE STORE

Word (*x*)	Semantic (*sem*)	Context (*c*)
Pen	a writing tool	Writing
Pen	a livestock's enclosure	Livestock
Pen	a portable enclosure for a baby	Play
Pen	a correctional institution	Institution
Pen	a female swan	Animal

5) **NLP Techniques**

The techniques of NLP that are involved in this work include syntactic processing and semantic processing. Syntactic processing is conducted by implementing a parser based on dynamic programming technique. The purpose of syntactic processing is to recognize the syntactic constituent in a sentence. Semantic processing is conducted by implementing λ reduction technique to attach semantics to the recognized constituents.

6) **Unambiguous Entities**

Unambiguous entities refer to the entities that have been identified its semantic based on the context of text. At this stage, one entity has one unique semantic as a result of the previous processes.

7) **Entity Knowledge-base**

It is a conceptual knowledge-base where all unambiguous extracted entities will be stored. The extracted entities are represented in a table form. The knowledgebase is assumed to be used by other types of text processing applications. Figure 2 illustrates the methodology of the technique based on the given framework. There are 7 steps. All steps are equally important. However, the "resolving ambiguous entity'' is considered as the heart of the technique. In this paper, only this step will be presented and discussed in details, in the following section.

B. Ambiguity Resolution

In this section, a proposed theory of how to resolve ambiguous entities using possibility theory and fuzzy sets is presented in details. Now, let us denote Ω as a set of lexical, F_c denotes a fuzzy set of Ω with subject to the context (C). Variable x is a lexical may be restricted by the fuzzy set F_c. We denote such a restriction as $\Pi(x, F_c)$, and call F_c the restricting fuzzy set of x. $\Pi(x, F_c)$ associates a possibility distribution with x. The possibility distribution function $\Pi^{x}_{F_c}(\mu)$ denotes the possibility for x to take value μ under the restriction of F_c. Numerically, the distribution function x under the restriction F_c is defined to be equal to the membership function of F_c, that is

$$\prod(x, F_C) = \mu Fc(X) \forall \mu \in \Omega \qquad (1)$$

Now let us consider the lexical semantics of the word 'pen'. Table II represents an assigned lexical context (C) to each of the semantics. The table is stored in the lexical knowledge store. Again, let us take the 'pen' as a lexical *x*, then lexical semantics of *x* can be formalized as

$$x = m_i, m_{i+1}, \ldots, m_j \quad (2)$$

where *mi* is the first semantic, and *mj* is the last semantic, and its membership function can be derived from Eq. 1 is

$$\prod(x, Fc(x) = (vi, vi+1, \ldots, v_j) \quad (3)$$

where *v* is a plausibility value, and it is context-dependent. When *x* is applied in a different context, it may take a different value. In this work, the *v* is assigned automatically and randomly by the processor. Figure 3 illustrates how the fuzzy values are assigned randomly to the semantics of lexical.

The most plausible value (ρ) of *x* is obtained by using maximum (*max*) operator of a fuzzy set. Thus

$$\rho = (v_i, v_{i+1}, \ldots, v_j) \quad (4)$$

Once ρ value has been calculated and presented, the most possible semantic can be attached to *x*. In this way, the lexical ambiguity can be resolved, consequently, the system is able to give the most accurate meaning or semantic of a given word. To apply the possibility theory to the technique, a *fuzzy semantics database* is created.

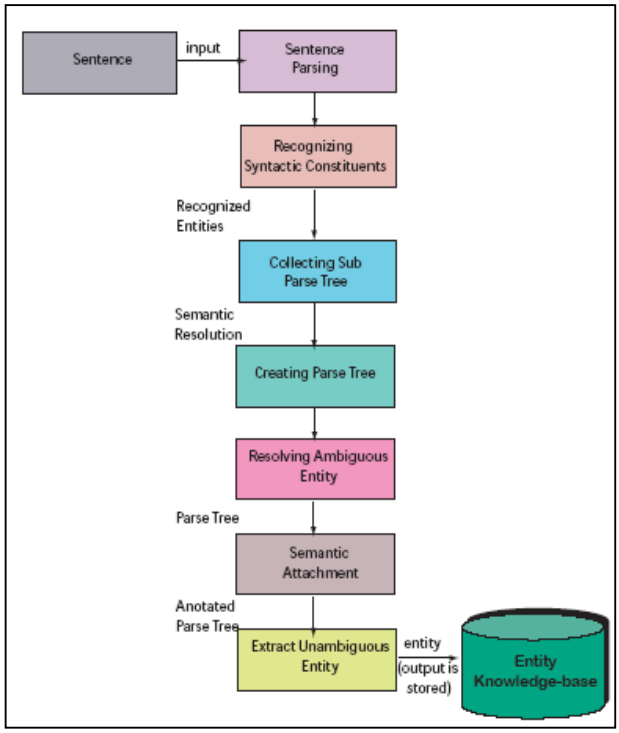

Figure 2. Methodology of the proposed technique

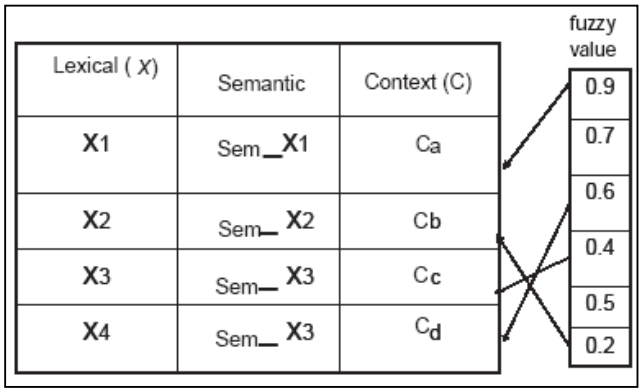

Figure 3. Fuzzy value is assigned to each semantic based on the context randomly

The fuzzy semantic database a table T contains three fields; lexical (x), semantic (sem_i) and semantic value (v). Conceptually the table T can be formalized as

$$T + \{((sem_i, v_i), (sem_{i+1}, v_{i+1}), (sem_j, v_j))\} \quad 5)$$

where sem_i denotes the meaning or semantic of the word x and v is a possibility value attach to it. The value v is in a range of (0, 1] based on the subject context. The values are generated and stored in the database manually by using human common sense. However, the process of assigning value to a lexical x is conducted dynamically and randomly.

TABLE III. FUZZY SEMANTIC DATABASE FOR "LIVESTOCK" CONTEXT

Word (x)	Semantic (sem)	Context (c)	*Grade* (v)
Pen	a writing tool	Writing	0.5
Pen	a livestock's enclosure	Livestock	0.9
Pen	a portable enclosure for a baby	Play	0.1
Pen	a correctional institution	Institution	0.2
Pen	a female swan	Animal	0.8

Figure 3 illustrates how values are assigned to semantics randomly. The result of value assignment will be a fuzzy semantic database as presented in Table III. Let us take *livestock* as a subject example. In the subject of *livestock,* a grade value of a word *pen* to have a semantic of *an enclosure for confining livestock* is 0.9 and *female swan* is 0.8. It is very unlikely for the word *pen* to be *a portable enclosure* in which babies may be left to play, thus the possibility value is 0.1. If the given subject is 'baby', its grade value might be 0.9. To resolve an ambiguous word *pen*, the most possible value ρ is calculated using the *max* fuzzy set operator. Using the example given in Table III, the ρ is calculated by using equation in 4. Using plausible value in the Table III, each v in
4 is replaced by v from the database, which can be represented

$$\rho = ((0.5, 0.9, 0.1, 0.2, 0.8)) \quad (6)$$

As the presented in Equation 6, the maximum value of all the plausible value in table T is 0.9, which 0.9 is taken the possible value for the word pen in the context of livestock. The semantic that attach the plausible value 0.9 is "An enclosure for confining livestock", consequently, its semantic is taken as the unique semantic of a word. When the most possible semantic is identified, the semantic attachment is conducted. Once semantic attachment process is completed, semantics for ambiguous entities have been resolved. At this stage, all recognized entities are not ambiguous anymore. The following step is to extract unambiguous entities. Syntactic constituents that are belongs to noun part-of-speech category are identified. The semantics of its constituent are extracted. The process of extracting entities is conducted by using a simple rule. The rule can be formulated as,

if the syntactic is noun, **then** extract its semantic.

The extracted semantic are then listed in a table form. The table is stored in the entity knowledge-base. The table consists of a list of sentences. Each sentence contains a set of semantics. The descriptions of the table can formalized in Equation 7 and 8

$$S_i = (sem_1, sem_2, sem_n) \quad (7)$$

where S denotes a sentence and i denotes a sentence index. As previously mentioned sem denotes a semantic.

$$Tab = (S_1, S_2, .., S_n) \quad (8)$$

where S_1, S^2, ...,S_n represent a semantic of entity in S_1, S^2, ...S_n consecutively.

IV. EXPERIMENT

Experiments have been carried out to evaluate the technique. This section presents, the procedures involved in conducting experiments, the environment of experiments, as well as the use of test case. The procedures of running experiments include define subject, create test cases, setup experiment environment and run the experiment. An experimental procedure is considered as a framework for testing the proposed technique. Each step in the methodology was tested. The step is considered successful if the output of the step can be used as an input to the following step.

A. Define Subjects

As previously mentioned, a subject with lexical knowledge is used to determine the context of the text. Thus, 20 subjects have been defined and stored. The defined subjects include institution, banking, business, human body, food, furniture, transportation, animal, sports, medical, plantation, device, documents, management, human baby, office, farming, distance, hardware tool and feelings. Words which are related to the defined subjects are stored. For example, the word 'bat' is stored for two subjects; animal and sport and the word 'chair' is stored for management and furniture subjects.

B. Develop Test Case

Test case is a well known method for testing software functions [21], [20], [7]. Our developed technique consists of several software functions because each step in the methodology was translated into a software function. Thus, the most suitable approach to evaluate the technique is by using test case approach. Therefore, 120 test cases have been developed. Each test case contains sentences in the range of 7 to 10, for different kinds of subjects. Most of the sentences in test case are extracted from the Internet as well as WordNet version 3.0. Table IV demonstrates sentences belong to human body subject. A word which may have more than one meaning in a different context is underlined.

TABLE IV. EXAMPLES OF SENTENCES FOR THE HUMAN-BODY SUBJECT

No	*Sentence*
1	He combed his hair.
2	Hair consists of layers of dead keratinized cells.
3	He stuck his head out the window.
4	His bare feet projected from his trousers.
5	His heart thumping wildly.
6	He has a cold in the nose

C. Setup Experiment Environment

The process of setting up experiments involves with storing words and its part-of-speech in a lexicon. In this work, the lexicon consists of 5000 words which belong to 8 part-of-speech categories. They are noun, verb, pronoun, preposition, adverb, adjectives, article (determiner) and auxiliary verbs. The same word may represent more than one syntactic constituent, for example, the word 'bat' has two types of constituent; verb and noun. Thus, in setting up the experiment environment, the word 'bat' is stored twice, first as a verb and second as a noun.

V. RESULT

Examples of test cases and sentences which have been used in the experiments are illustrated in figure 4, 5, 6 and 7. Each figure consists of a 4 column table, which representing test case number, test conditions, test sentences, and results given by the system. In this section, a 'system' is referred to the implemented technique. Results of each test case were obtained based on the test conditions being used. The following is a list of the test conditions:

- A test case contains a sentence that has an ambiguous and unambiguous entity.
- A test case contains a sentence that has a comma.
- A test case contains a sentence that has a conjunction word.

- A user may enter incorrect subject.
- A user enters correct subject.

Test Case	Test Condition	Sentence	Result
1	Mix ambiguous entity and unambiguous entity Use correct subject	1. I cut my hair 2. The bat slept 3. The chair is in the office 4. I went to a bank yesterday 5. He is writing using a new pen 6. He killed the bat with a new gun 7. I bought a new chair	AE = 7 AEE = 7 CAEE =7 PCEE =7
2	Mix ambiguous entity and unambiguous entity A sentence consists of a comma Use incorrect subject	1. The bank opened a branch in Kuching 2. He has a nose for good deals * 3. He bought a new bat yesterday at the sport shop 4. Yes, he is the chair* 5. The boy is writing using a pen 6. He has a cold in the nose 7. Who is the chair of this department?* 8. Sugar is expensive in Jordan 9.Her blood contain lots of sugar* 10. We sat on the bank*	AE = 11 AEE = 9 CAEE =8 PCEE =11
3	Mix ambiguous and unambiguous entity Use incorrect subject	1. Pizza has too much fat 2. The gardener planted a bed of roses 3. She dislike fatness in her body 4. He sat on the bed 5. The stock market fell into a new low 6. The established a cap for prices 7.She has an eye for a fresh talent* 8. He tried to catch her eye* 9. Where did you buy the bat? 10.The whole town cheered for the team	AE = 10 AEE = 10 CAEE =8 PCEE =10

Figure 4. Results of test case 1, test case 2 and test case 3

The results were calculated based on the number of ambiguous entity (AE) in a test case, the number of ambiguous entity that was successfully extracted by the system (AEE), the number of correct semantic attachment for the ambiguous extracted entity by the system (CAEE) and the number of predicted correct semantic attachment for ambiguous entity (PCEE). Obtained results indicate that the number of AE is same as the number of PCEE. This scenario happens because the focus of the testing is for ambiguous entities only.

Test Case	Test Condition	Sentence	Result
4	Use incorrect subject A sentence consists of a conjunction word	1. His heart thumping wildly 2. The satellite dish is not expensive 3. She cooked a nice dish yesterday 4. The flight is late 5. The meeting has been cancelled 6. Ahmad and Armin do not pay the loan* 7. When he returned to work he met many new faces* 8. He answered with the truth but they would not believe *	AE = 8 AEE = 6 CAEE =5 PCEE =8
5	Mix ambiguous and unambiguous entity Use incorrect subject	1. He stuck his head out of window 2. The thread would not go thorough the eye 3. They sell 200 head of cattle 4. He admired her long graceful neck 5. The bottle had a wide neck* 6. He has plenty of brains but no common sense * 7. I could not get his words out of my head 8. His mind wandered 9. The body of the car is badly rusted 10. She does not have a heart to betray her*	AE = 10 AEE = 10 CAEE =7 PCEE =10
6	Mix ambiguous and unambiguous entity Use incorrect subject A sentence consists of a conjunction word	1. The spirit is willing but the flesh is weak * 2. My body is weak 3. The body of the message is short* 4. It came to my mind 5. He has no stomach for a fight 6. She has a big stomach 7. The whole body filled out the auditorium 8. The arm of the record player 9. I have broken my arms 10. He reads to improve his mind*	AE = 10 AEE = 9 CAEE =7 PCEE =10

Figure 5. Results of test case 4, test case 5 and test case 6

Unambiguous entity, a sentence that has a comma or a conjunction word, and a user enters an incorrect subject are used as negative conditions. While positive conditions include a sentence has ambiguous entity and a user enters a correct subject. Positive conditions are used as a boundary analysis for what the system is supposed to do and negative conditions are used as a boundary analysis for what the system is not supposed to do. The obtained results indicate that a test condition of mixing ambiguous and unambiguous entities in a sentence does not cause any problem to the system. Results of test case 1, 8, and 10 in Figure 4, 5, 6 consecutively, illustrate that the system is able to extract all ambiguous entities correctly whenever a human user uses correct subjects as defined in the system (Note that AEE = CAEE). However, the system is not able to extract all ambiguous entities correctly whenever a human user uses incorrect subject as illustrated by test case 2, 3, 4, 5, 6, 7, 9, 11, and 12.

Another interesting scenario occurs whenever a sentence consists of a comma or a conjunction word. The obtained results showed that the number of extracted ambiguous ambiguity is not same as the number of ambiguous entities in the test case (Note that AEE _= AE). This scenario happens because when a sentence has a comma or a conjunction word, a sentence cannot be successfully parsed by the system. Consequently, syntactic constituents in a sentence cannot be recognized. The obtained results show that the system is able to recognize 1120 ambiguous entities and successfully attach correct semantics to 900 ambiguous entities.

Test Case	Test Condition	Sentence	Result
7	Mix ambiguous and unambiguous entity A sentence consists of a comma Use incorrect subject	1. Get out of my way, boy* 2. His boy is taller than him 3. Exercise give him a good stomach for dinner 4. Do not pay him any mind 5. How our minds work? 6 You have not got the heart for baseball 7. Where is your taxi? 8.The color of the blood is red 9. We need young blood to do the job* 10.He is the head of the operation	AE = 9 AEE = 8 CAEE =7 PCEE =9
8	Mix ambiguous and unambiguous entity Use correct subject	1. My horse lost the race by a nose 2. They bought a new bat at the sport shop 3. He bought a new play pen for his baby yesterday 4. The farmers raise pens in their lakes 5. Tickets are five dollars per head 6. The boy is writing using a pen 7. His bare feet projected from his trousers 8. He has a cold in the nose 9. Sugar is not good for a body 10.We buy a set of chairs	AE = 11 AEE = 11 CAEE =11 PCEE =11
9	Mix ambiguous and unambiguous entity Use incorrect subject A sentence consists of a conjunction word	1. She washes dishes in the sink 2. I saw through his little game from the start* 3. The body of faculty and students at a university* 4. He washed his face 5. She has broken her arms but she does not give up* 6. His new bat has broken 7. She cuts vegetables with a knife 8. I have pain in my nose 9. That room is 50 feet long	AE = 9 AEE = 8 CAEE =7 PCEE =9

Figure 6. Results of test case 7, test case 8 and test case 9

Test Case	Test Condition	Sentence	Result
10	Mix ambiguous and unambiguous entity Use correct subjects	1. A mouse takes more space than a trackball 2. The cat caught a small mouse 3. She holds a new glass 4. He grows plants in his garden 5. He has written many scientific papers 6. Papers are made of wood 7. The notion of an office running without paper is absurd 8. I reserved a table at my favorite restaurant 9. See table 4 in page 10	AE = 9 AEE = 9 CAEE =9 PCEE =9
11	Mix ambiguous and unambiguous entity Use incorrect subject	1. His students followed him like sheep* 2. The farmers keep chickens in a pen 3. The face of the city is changing * 4. The plane was delayed 5. The cabinetmaker used a plane for the finish work 6. We gave them a set of dishes for a wedding present 7. She prepared a special dish for dinner* 8. The capital allowance amount is computed on a straight-line basis 9. The national carrier managed to hold up 10.He stopped the car and turned off the lights	AE = 11 AEE = 11 CAEE= 8 PCEE =11
12	Mix ambiguous and unambiguous entity A sentence consists of a conjunction word Use incorrect subject	1. He is sitting in the room like a mouse* 2. She washes her palm thoroughly 3. Farmers in Indonesia started to grow palm 4. I do not know she is chicken* 5. Gardener always cut unnecessary branch 6. Her fingers were long and thin * 7. The price of nails is expensive 8. Nail is a former unit of length for cloth*	AE = 7 AEE = 7 CAEE =6 PCEE =7

Figure 7. Results of test case 10, test case11, and test case 12

$$\Pr ecision = \frac{total\_CAEE}{total\_AEE} \qquad (9)$$

Equation 10 illustrates that the system achieves 85.7% of accuracy based on the test cases used. Although the precision metric is enough to evaluate the system, recall metric is also used to analyze the obtained results against the expectation of the system.

$$\Pr ecision = \frac{900}{1050} = 0.857 \qquad (10)$$

Recall is defined as *the total of correct entity extracted by the system* divided by *the total number of expected correct entity extracted by the system* (see Equation 11).

$$\mathrm{Re}\,call = \frac{total\_CAEE}{total\_PCEE} \qquad (11)$$

$$\mathrm{Re}\,call = \frac{900}{1120} = 0.803 \qquad (12)$$

VI. Discussion

Central interest of this paper is to seek a new technique for disambiguating ambiguous entities. The new technique is classified as an unsupervised system. The number of test cases and sentences used in assessing the technique is considered reasonable enough because it can cover positive and negative conditions to demonstrate the system's behavior. Again, the aim of the testing is to evaluate each step of the technique. Furthermore, [51] reported that many word sense disambiguation (WSD) systems are based on limited number of words. They also reported that [53] and [47] work published results for 12 words, while [30] and [8] published results for just one word. The extensive survey (69 pages) published by [38] concluded that all of the WSD systems are evaluated independently, in which no comparison is made between one system to another.

Based on the above justifications, results which are gathered from the experiments have been analyzed. Precision and recall are used as measurement metrics. From the result analysis, it can be concluded that, the proposed technique is able to resolve ambiguous entities within the positive condition tests. The achieved accuracy and recall percentage rate depends on the test conditions being used as well the capability of the parser or syntactic recognizer. The results also reveal that the percentage rate of recall is less than the percentage rate of precision. This scenario occurs when a sentence cannot be parsed successfully. This happens when a sentence has a comma or a conjunction word or the structure of the sentence cannot be covered by the existing grammar rules. Consequently, a correct entity cannot be recognized and therefore, the system cannot attach a correct semantic to the entity. This scenario demonstrates that the capability of the parser or syntactic recognizer also play an important role in an entity extraction. However, to build a robust parser is not the focus of this paper. Up to date, a robust parser that can parse all kind of sentence is still not available. Thus, a parsing research area is still open to be explored by researchers. Although the work in this paper is similar to WSD and NER work, they are not the same. There are two reasons for this. The first is WSD systems focus on ambiguous words which are not limited to noun part-of-speech only. Other part-of-speech such as a verb and an adjective are also been taken care of. For example, the word 'bat' would be resolved into a noun sense and a verb sense, and then a correct sense is tagged to the word. In this work, the former problem has been resolved during the parsing process, where the syntactic recognizer is able to recognize the word 'bat' as a noun constituent or a verb constituent. If it is a noun constituent, then the process of resolving ambiguous entity is conducted, otherwise it is not a problem to the system. Secondly, in NER, the system is supposed to recognize and extract the named entity such as the name of a person, location, organization, date and time. The ambiguity problem is not a focus because the system does have to attach a correct semantic to the extracted named entity. It is a process of filling the predefined template with correct entities [14], [12].

VII. Conclusion

This paper has presented a new technique for resolving ambiguity in entity extraction. The technique which is obtained by combining context knowledge and fuzzy approach is integrated into NLP. The obtained results show that fuzzy approach and context knowledge can be used as tools in resolving lexical ambiguity problem. The significant contribution of this research work is a new technique which can be used in text pre-processing in text mining applications such as machine translation, automated text summarization, automated text categorization and many more.

References

[1] E. Agirre and P. Edmonds. Introduction. In E. Agirre and P. Edmonds,editors, Word Sense Disambiguation: Algorithms and Applications, pages 1–28. Springer Verlag, New York, 2007.

[2] E. Agirre and D. Martinez. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the Semantic* Annotation And Intelligent Annotation workshop organized by COLING *Luxembourg 2000*, pages 11–19, 2000.

[3] E. Agirre and M. Stevenson. *Knowledge sources for WSD*, pages 217 -251. Springer Verlag, New York, 2007.

[4] J. Allen. *Natural Language Understanding*. The Benjamin/Cummings Publishing Company, United States of America, 1988.

[5] K. L. Baker, A. M. Franz, E. M. Franz, and P. W. Jordan. Coping with ambiguity in knowledge-based natural language analysis. In *Proceedings* of FLAIRS, 1994

[6] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In Proceedings of the 18th international joint conference on Artificial intelligence, pages 805–810, 2003.

[7] D. L. Bird and C. Munoz. Automatic generation of random self-checking test cases. *IBM System Journal*, 22(3):229–245, 1983.

[8] R. Bruce and J. Weibe. Word-sense disambiguation using decomposable models. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL), pages 139–145, 1994.

[9] C.-H. Chang, M. Kayed, M. R. Girgis, and K. Shaalan. A survey of web information extraction systems. *IEEE Transactions of Knowledge and Data Engineering*, 18(10):1411–1428, 2006.

[10] C. M. Chiara, F. Fernando, and G. Patrizia. Ambiguity detection in multimodal systems. In Proceedings of the Working Conference on *Advanced visual interfaces*, pages 331–334, New York, NY, USA, 2008. ACM.

[11] C.-T. Chu, Y.-H. Sung, Z. Yuan, and D. Jurafsky. Detection of word fragments in mandarin telephone conversation. In *International* Conference on Spoken Language Processing, 2006.

[12] J. Cowie and Y. Wilks. *Information extraction*. New York, 2000.

[13] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Unsupervised named-entity extraction from the Web: an experimental study. *Artificial Intelligence*, 165(1):91–134, 2005.

[14] R. Feldman and J. Sanger. The text mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, United State of America, 2007.

[15] M. Fleischman and E. Hovy. Fine grained classification of named entities. In Proceeding of the 19th International Conference on Computational Linguistics (COLING), 2002.

[16] T. Grenager, D. Klein, and C. D. Manning. Unsupervised learning of field segmentation models for information extraction. In *Proceedings of* the 43rd Annual Meeting on Association for Computational Linguistics, pages 371–378, 2005.

[17] R. Grishman. Information extraction: Techniques and challenges. In Proceedings of the SCIE, pages 207–220, 1997.

[18] R. Grishman. NLP: An information extraction perspective. In *Proceedings* of the Recent Advances in Natural Language Processing (RANLP *2005)*, 2005.

[19] R. Hale. Text mining: Getting more value from literature resources. *Drug Discovery Today*, 10(6):377–379, 2005.

[20] C.-Y. Huang, J.-R. Chang, and Y.-H. Chang. Design and analysis of GUI test-case prioritization using weight-based methods. *Journal System Software*, 83(4):646–659, 2010.

[21] D. Jeffrey and N. Gupta. Experiments with test case prioritization using relevant slices. *Journal System Software*, 81(2):196–221, 2008.

[22] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of the 10th International Conference on Research in Computational Linguistics, September 1997.

[23] D. Jurafsky and J. H. Martin. Speech and Language Processing: An introduction to Natural language Processing, Computational Linguistics *and Speech Recognition*. Prentice Hall, United States of America, 2009.

[24] D. Jurafsky, R. Ranganath, and D. McFarland. Extracting social meaning: identifying interactional style in spoken conversation. In *Proceedings* of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational *Linguistics (NAACL '09)*, pages 638–646, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[25] S. Jusoh and H. M. Alfawareh. Natural language interface for online sales. In Proceedings of the International Conference on Intelligent an *Advanced System (ICIAS2007)*, pages 224–228, Malaysia, November 2007. IEEE.

[26] E. Kamsties. Surfacing Ambiguity in Natural Language Requirements. Phd thesis, Fraunhofer- Institue fr Experimentelles Software Engineering, 2001.

[27] H. Karanikas, C. Tjortjis, and B. Theodoulidis. An approach to text mining using information extraction. In *Proceedings of Workshop of* Knowledge Management: Theory and Applications in Principles of Data Mining and Knowledge Discovery 4th European Conference, 2000.

[28] C. Karat, J. Vergo, and D. Nahamoo. Conversational interface technologies. In J. A. Jacko and A. Sears, editors, *The Human-Computer Interaction Handbook*, pages 169–186. Lawrence Erlbaum Associates, 2003.

[29] G. J. Klir and B. Yuan. Fuzzy Sets and Fuzzy Logic: Theory and *Applications*. Prentice Hall, United States of America, 1995.

[30] C. Leacock, G. Towell, and E. Voorhees. Corpus-based statistical sense resolution. In Proceedings of the ARPA Workshop on Human Language *Technology*, pages 260–265, 1993.

[31] H. Liu, S. B. Johnson, and C. Friedman. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. Journal of the American Medical Informatics Associations *(JAMIA)*, 9:621–636, 2002.

[32] R. Malik. CONAN: Text Mining in Biomedical domain. Phd thesis, Utrecht University, Austria, 2006.

[33] D. McCarthy, J. Carroll, and J. Preiss. Disambiguating noun and verb senses using automatically acquired selectional preferences. In Proceedings of the SENSEVAL-2 Workshop at the European Chapter *ACL*, pages 119–122, Toulouse, France, 2001.

[34] R. Mihalcea and D. Moldovan. A highly accurate bootstrapping algorithm for word sense disambiguation. *International Journal of* Artificial Intelligence Tools, 10(1-2):5–21, 2001.

[35] R. Milne. Parsing against lexical ambiguity. In *Proceedings of the 8th* conference on Computational linguistics, pages 350–353, Morristown, NJ, USA, 1980.

[36] R. J. Mooney and R. Bunescu. Mining knowledge from text using information extraction. ACM SIGKDD Explorations Newsletter, 7(1):3 10, 2005.

[37] R. Navigli. A structural approach to the automatic adjudication of word sense disagreements. Journal of Natural Language Engineering, 14(4):293–310, 2008.

[38] R. Navigli. Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69, 2009.

[39] R. Navigli and P. Velardi. Automatic adaptation of WordNet to domains. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), 2002.

[40] R. Navigli and P. Velardi. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 27(7):1075–1088, 2005.

[41] C. Nedellec and A. Nazarenko. Ontologies and information extraction: A necessary symbiosis. In P. Buitelaar, P. Comiano, and B. Magnin, editors, Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press Publication, 2005.

[42] F. Portet, E. Reiter, J. Hunter, and S. Sripada. Automatic generation of textual summaries from neonatal intensive care data riccardo bellazzi. In A. Abu-Hanna and J. Hunter, editors, *11th Conference on Artificial* Intelligence in Medicine (AIME 07), pages 227–236. Springer-Verlag, 2007.

[43] D. Ramage, E. Rosen, J. Chuang, C. D. Manning, and D. A. McFarland. Topic modeling for the social sciences. In *Workshop on Applications* for Topic Models: Text and Beyond (NIPS 2009), Whistler, Canada, December 2009.

[44] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the International Joint Conference Artificial Intelligence (IJCAI), pages 448–453, 1995.

[45] G. Ritchie and H. Thompson. Natural language processing. In: Artificial Intelligence: Tools, Techniques, and Applications. Harper and Row, 1984.

[46] R. Schank and R. Abelson. Scripts, Plans, Goals, and Understanding. Lawrence Erlbaum, Hillsdale, N.J, 1977.

[47] H. Schutze. Dimensions of meaning. In *Proceedings of Supercomputing' 92*, pages 787–796, 1992.

[48] T. Sekimizu, H. Park, and J. Tsuji. Identifying the interactions between genes and gene products based on frequently seen verbs in medline abstract. Universal Academy Press, Tokyo Japan, 1998.

[49] S. Sekine and C. Nobata. Definition, dictionaries and tagger for extended named entity. In Proceedings of the Conference on Language Resources and Evaluation, 2004.

[50] N. Singh. The use of syntactic structure in relationship extraction Master's thesis, MIT, 2004.

[51] M. Stevenson and Y. Wilks. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistic*, 27(3):321–349, 2001.

[52] Y. Wilks. A preferential pattern-seeking semantics for natural language inference. *Artificial Intelligence*, 6:53–74, 1978.

[53] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics(ACL-95), pages 189–196, 1995.

[54] L. A. Zadeh. Pruf- a meaning representation language for natural languages. International Journal of Man-Machine Studies, 10:395–459, 1978.

[55] S. Zhao and D. Lin. A nearest-neighbor method for resolving Ppattachment ambiguity. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP-04), pages 545– 554, 2004.

AUTHORS PROFILE

Hejab M. Alfawareh is an assistant professor of the Department of Information System in the Faculty of Science & Information Technology at Zarqa University, Jordan. He earned his Bachelor and Master degrees from Ukraine and earned his PhD degree in Information Technology with specialization in Artificial Intelligence, from Malaysia.His research interests include information extraction, natural language processing, computer networks, and fuzzy approach.

Shaidah Jusoh is an associate professor of the Department of Computer Science in the Faculty of Science & Information Technology at Zarqa University, Jordan. She **e**arned degrees of Master in Computer Science and PhD in Engineering System & Computing from University of Guelph, Canada. Her research interests include, an automated text summarization, question-answering system, online social network, data mining, and information extraction. She has published more than 50 articles in well known international journals and proceedings.