# Extracting Semantics for Information Extraction

**Hejab M. Al Fawareh[a], Shaidah Jusoh[b]**

[a,b.]*College of Arts and Sciences*
*Universiti Utara Malaysia, 06010 Sintok, Kedah*
*Tel : 04-9284701, Fax : 04-9284753*
*E-mail :alfawareh@gmail.com, shaidah@uum.edu.my,*

**ABSTRACT**

*Text documents are one of the means to store information. These documents can be found on personal desktop computers, intranets and in the Web. Thus the valuable knowledge is embedded in an unstructured form. Having an automated system that can extract information from the texts is very desirable. However, the major challenging issue in developing such an automated system is a natural language is not free from ambiguity and uncertainty problems. Thus semantic extraction remains a challenging task to researchers in this area. In this paper, a new framework to extract semantics for information extraction is proposed, where possibility theory, fuzzy sets, and knowledge about the subject and preceding sentence have been used as the key in resolving the ambiguity and uncertainty problems.*

**Keywords**
*Semantic Extraction, Information Extraction, Possibility Theory, Fuzzy Sets*

## 1.0 INTRODUCTION

Nowadays, the Web is considered as the world's largest repository of knowledge, and it is being constantly augmented and maintained by millions of people around the world. However, it is not in the form of a database from which records and fields are easily manipulated and understood by computers, but in natural language texts which are intended for human reading. In spite of the promise of the semantic web, the use of English language and other natural language texts will continue to be a major medium for communication, knowledge accumulation, information distribution on the Web, emails, reports, memos, blogs and etc. (McCallum, 2005).

People want to extract useful information from the texts documents quickly at a low cost. Text mining is a new area which focuses on the use of automated methods for exploiting the enormous amount of knowledge available in text documents. Text mining, sometimes alternately referred to as text data mining, refers generally to the process of deriving high quality information from texts (AlFawareh *et al.*, 2008). Typical text mining tasks include text categorization, text clustering, concept/entity and fact extraction, and production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (Redfearn, 2006).

When dealing with natural language texts, the most critical problem is ambiguity and uncertainty issues. Automated information extraction (IE) system should be able to extract correct semantics from texts. Thus the ambiguity and uncertainty issues should be resolved. In this research work, we propose a new framework of semantic extraction for IE by using *the knowledge of subject* and *relevant preceding sentence*. This paper is organized as follows. Section 2.0 will discuss information extraction; section 3.0 will present a proposed framework. The implementation and result analysis are presented in section 4.0. Section 5.0 concludes the paper.

## 2.0 RELATED RESEARCH

IE involves directly with text mining process by extracting useful information from the texts and stores them into a structured database. Then data mining techniques can be applied to the data for discovering new knowledge. According to Grishman (1997), 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. Traditionally in IE, natural language texts are mapped to predefined, structured representation, or templates, which, when filled, represent an extract of key information from the original text (Karanikas *et al.*, 2000).

In IE, there are two levels of extractions; entity extraction and fact extractions. Extracting entity/concepts from the texts require a person to read them. Fact extraction is a process of spreading out the facts from entities. This is very time consuming. It can become a challenging task if the person does not have enough background related to the texts. Having an automated system that can extract required information from the texts is becoming an urgent need. However, this desire is not easy to achieve. Natural language texts are not free from the ambiguity problems. It is not only many words may refer to one meaning and one word may have more than one meaning, but also a structure of the sentence can be interpreted into more than one meaning.

On the other hand Singh (2004) and Hale (2005) addressed information extraction is based on understanding of the structure and meaning of the natural language in which documents are written and the goal of information extraction is to accumulate semantic information from text. Technically extracting information from texts requires lexical knowledge, grammars describing the specific syntax of the texts to be analyzed as well as semantic (Nedellec and Nazarenko, 2005).

Today, most of the IE systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. However, the growing need for IE application to domains such as functional genomics that require more text understanding. Named entity recognition describes the identification of entities in free text. For example, in biomedical domain, entities would be gene, protein names and drugs. NER often forms the starting point in a text mining system, meaning that when the correct entities are identified, the search for patterns and relations between entities can begin. Malik (2006) also claim that one of the major problems in NER is ambiguous protein names; one protein name may refer to multiple gene products.

Although Liu *et al.* (2003) have put effort to resolve ambiguous terms using sense-tagged corpora and UMLS, the ambiguity is still the major "world problem" (Malik, 2006). In fact Liu and co-workers work focus only on biomedical terms only. Recognizing and classifying named entities in texts require knowledge on the domain entities. List entities are used to tag text entities, with the relevant semantic information; however exact character strings are often not reliable enough for precise entity identification (Nedellec and Nazarenko, 2005).

Recent applications in information extraction include apartment rental ads (Soderland, 1999), job announcements (Calliff and Mooney, 1999), geographic web documents (Etizioni *et al.*, 2004), government reports (Pinto *et al.*, 2003) and medical abstracts (Malik, 2006). Yangarber et. al (2005) point out that much published work on IE reports on closed experiments; systems are built and evaluations are conducted based on carefully annotated training and test corpora. Although IE has been implemented for varieties of applications as mentioned above, up to date, semantic extraction has not yet been fully explored for IE.

## 3.0 PROPOSED FRAMEWORK

Our proposed framework solves the ambiguity and uncertainty problems in semantic extraction for IE at two levels of extraction. The first is at the entity extraction level and the second is at the fact extraction level as shown in Figure 3.1. The whole process of extracting entity and facts from texts can be condensed into 3 steps as illustrated in Figure 3.1

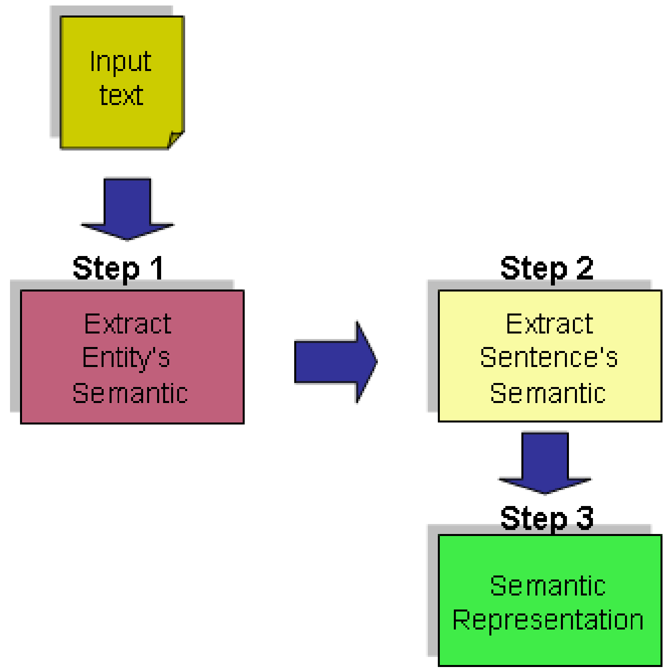


*Figure 1:* The steps for semantic extraction

### 3.1 Step 1

In this step, the text input is segmented into sentences. Each sentence will be processed syntactically to recognize its part of speech. The word that is belong to a verb or a noun part-of-speech category is defined as an entity. Let us consider the following sentences as examples:

- I put the baby in the <u>pen</u>
- She <u>runs</u> the company

From syntactic processing, the system would be able to determine that the word pen is belong to part-of-speech for noun category. The syntactic processor also can determine that "runs" is a verb. However, when the system needs to extract the semantic of the word, the system would face ambiguity and uncertainty problem. For example, a word `pen' can be interpreted as a writing tool, or an enclosure, in which babies may be left to play. While the word `run' can be interpreted as an activity of controlling or as a physical action. In information extraction, semantic of the texts should be correctly interpreted.

To resolve the problem we have applied subject context knowledge during the semantic processing. Figure 3.2 illustrates the process.

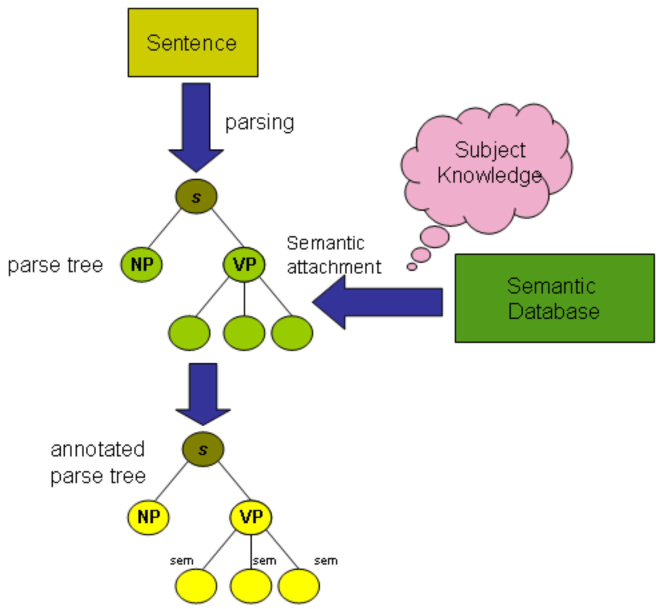


*Figure.2*: The process of Step 1

As previously mentioned, the structure of a sentence (parse tree) is obtained through a parsing/syntactic process. Using the possibility theory, we assign the possibility value to the meaning of the words. The value is determined by the context knowledge. Let us consider a pen as a word ($w$) and its meanings; a tool for writing ($m_1$) and an enclosure ($m_2$). The possibility (ρ) of w = m1 or w= m2, is determined by context knowledge (CK), which can be formulated as follows

$$w = (m_1, m_2, m_3, \ldots m_n) \qquad (1)$$

where m1, …mn, represent the possible meaning of the word w, and n is a finite number of the meaning.

$$\rho = (\rho_1 = m_1, \rho_2 = m_2,\ \rho_3 = m_3,\ \rho_n = m_n) \qquad (2)$$

The possible meanings of w is represented by $\rho_1$, $\rho_2$... $\rho_n$. The value of $\rho_1$, $\rho_2$... $\rho_n$ is decided based on the CK as represented in the Table 1.

Table 1: *A semantic database for "baby" context.*

| Word (w) | Semantic (m) | Possibility Value (ρ) |
|---|---|---|
| Pen | A writing tool with a point from which ink flows | 0.5 |
| Pen | An enclosure for confining livestock | 0.1 |
| Pen | An enclosure in which babies may be left to play | 0.9 |
| Pen | A correctional institutions for those convicted crime | 0.2 |
| Pen | Female swan | 0.4 |

In Table 1, the context of the word pen is "baby". In this work, fuzzy operator *max* is used to select the most possible meaning of the pen as formulated in Eq. 3

$$\rho = \max(\rho_1, \rho_2, \rho_3, \ldots, m_n) \qquad (3)$$

Thus, by applying Eq. (3), the syntactic processor is able to decide the most possible meaning of the word `pen', which is an enclosure in which babies are left to play. Therefore, if the subject knowledge is "writing" the values of the possibility in Table 1 would be different. Once the ambiguity and uncertainty problems, a correct semantic is attached to the parse tree. The annotated parse tree would be used for the process in the step 2.

### 3.2 Step 2

In Step 2, annotated parse tree is used to determine the semantic meaning of the sentence. Let us consider the sentence "I put the baby in the pen". Although, step 1 has resolved that the ambiguity problem for the word pen, during the parsing process, the syntactic processor would also generate more than one parse tree. This happens because of the ambiguity in the grammar itself. The sentence can be parsed in two ways; the first parse tree is parsed through production grammar rules in 1, and the second parse tree through production grammar rules in 2, as illustrated below.

1. s → Pronoun, VP
   VP → Verb,NP,PP  parse tree 1

2. s → Pronoun, VP
   VP → Verb,NP  parse tree 2
   NP → Det, Noun, PP

When the sentence can be parsed in two ways, there will be two possible meanings of the sentence. The first parsing could be interpreted as the "the person put the baby who is located at some place into the pen" and the second parsing could be interpreted as "the baby is already in the pen, and the person put him/her into some place". To extract semantic from the sentence, the processor should be able to determine the most possible meaning.

To resolve the problem, the processor refers to the previous related sentence and uses its semantic to determine most possible meaning of the current sentence. As example, the preceding sentence of the sentence "I put the baby in the pen" is "A baby is left alone on the floor". By using the knowledge about the most relevant preceding sentence, a possible value ($\sigma$) is attached to the derived production rules. Thus the production rule of grammar can be represented as $\alpha \xrightarrow{\sigma} \beta$ where $\sigma$ is a *plausibility function* in each grammar rule, and $\sigma \in [0,1]$ indicates the plausibility for substituting $\alpha$ with $\beta$ in a parsing process. A string $S$ of symbols in $V_T$ is said to be in the language $L$ ($G$) if and only if $s \to S$, i.e. $S$ is derivable from $s$. When $Tr$ is a parse tree generating $S$, the plausibility of $Tr$ is

$$\min\{\mu(s \to \alpha_1),\ldots,(\alpha_n \to S)\} > 0 \qquad (4)$$

where $s \to \alpha_1, \alpha_1 \to \alpha_2, ..., \alpha_m \to S$ is the derivation chain from which $Tr$ is constructed, and $\mu(\alpha_i \to \alpha_{i+1})$ is the non-zero $\sigma_{(i+1)}$. The restricting fuzzy set $F_s$ is defined as

$$Fs = \{Tr\} \qquad (5)$$

and its membership function is

$$\mu F_s(Tr) = \begin{cases} \min\{\mu(s \to \alpha_1),\ldots,(\alpha_m \to S)\} & \text{if } s \to_{Tr} S \\ 0 & \text{otherwise} \end{cases}$$

where $\to_{Tr}$ is the chain $s \to \alpha_1, \alpha_1 \to \alpha_2, \ldots, \alpha_m \to S$ from which $Tr$ is constructed. When a sentence is ambiguous, the fuzzy max operator is used to select the most possible parse tree, which is formulated in Eq. (6)

$$\mu F_s(G_{Tr}) = \{\max(Tr_1,\ldots,Tr_n)\} \qquad (6)$$

Semantically, the sentence "I put the baby in the pen" is resolved to the meaning "the person put the baby who is located at some where into a pen". For further computation, predicate calculus is used for semantic representations.

## *4.0* IMPLEMENTATION AND RESULTS

The proposed framework has been implemented in C language on Linux Operating system. Dynamic programming technique has been used to create a parser for syntactic processing, where Earley Algorithm (Earley, 1970) has been applied. The semantic attachment has been conducted by using lambda reduction technique (Jurafsky & Martin, 2000). In this work, seventy fuzzy grammar rules have been used. Fifteen data sets have been used for the framework. Each data set consists of ambiguous and unambiguous sentences. Each sentence may contain ambiguous and unambiguous words. The length of data set is between five to seven sentences. The process is conducted at a constituent and sentence level. Table 2 presents some of the obtained results. The extracted semantics after ambiguity resolution have been presented in a predicate calculus. As we can see from the Table 2, some sentences have only one level ambiguity and some have two levels of ambiguity. For example, sentence no. 1 has two levels of ambiguity. The first level is constituent level which has been resolved using the technique in Step 1. The second level is sentence level, which the fact ambiguity has been resolved using the technique discussed in Step2. The obtained results indicate the proposed techniques in Step 1 and Step 2 are successful.

Table 2: *Results of the semantics extractions*

| No. | Sentence | Ambiguity Level | Extracted Semantics after Ambiguity Resolution |
|---|---|---|---|
| 1 | I put the baby in the pen | constituent & sentence | put (baby, play_pen) |
| 2 | She runs the company | Constituent | manage(person, company) |
| 3 | The boy washes the plate in the sink | Sentence | wash(plate, sink) |
| 4 | She combs hair with a comb | constituent & sentence | comb(person, human-hair) |
| 5 | The chair is in the office | Constituent | is_in(chairman, office) |
| 6 | I bought a new chair yesterday | Constituent | bought(speaker, chair) |
| 7 | There are twenty heads in the room | Constituent | are_in(persons, room) |

## 5.0 SUMMARY

This paper proposes a new framework for extracting semantics from texts. The novelty of this framework is the knowledge of about subject and the most relevant preceding sentence have been used to resolve ambiguity in extracting semantics for information extraction. Possibility theory and fuzzy sets have been used to extract the most possible semantics from the texts based on the knowledge about subject and preceding sentence. Experimental results indicate that the proposed framework is successful.

## REFERENCES


AlFawareh, H. M., S. Jusoh and W. R. S. Osman (2008). Ambiguity in text mining. In: Proceedings of the International Conference on Computer and Communication Engineering (ICCCE'08). pp. 1172–1176.

Calliff, M.E. and R.J. Mooney (1999). Relational learning of pattern-match rules for information extraction. In: *Proceedings of the Sixteenth National Conference on Artificial Intelligence*. pp. 328–334.

Earley, J. (1970), An efficient context-free parsing algorithm. Communication of the ACM Vol. 13, No. 2, pp 94-102

Etizioni, O., Cafarella M, M. Downey, S. Kok, A-M. Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates (2004). Web-scale information extraction in know it all. In: *Proceedings of the Thirteenth International World Wide Web Conference*

Grishman, R. (1997). Information extraction: Techniques and challenges. In: *SCIE*. pp. 10–27.

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

Jurafsky, D. & Martin, J. H. (2000), *Speech and Language Processing,* United States of America, Prentice-Hall.

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

Liu, Q., H.Yu, X.Cheng, and S. Bai, (2003), Chinese Named Entity Recognition Using Role Model, *Computational Linguistics and Chinese Language Processing,* Vol. 8, No. 2, pp 1-31

McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. Queue 3(9), 48–57.

Malik, R. (2006). CONAN: Text Mining in Biomedical domain. PhD thesis. Utrecht University, Austria

Nedellec, C. and A. Nazarenko (2005). Ontologies and information extraction: A necessary symbiosis. In: *Ontology Learning from Text: Methods, Evaluation and Applications* (P. Buitelaar, P. Comiano and B. Magnini, Eds.). IOS Press Publication

Redfearn, J. (2006). Text mining. JISC pp. 1–2.

Singh, N. (2004). The use of syntactic structure in relationship extraction. Master's thesis. Department of Electrical Engineering and Computer Science, MIT.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine Learning* **34**, 233–272.

Yangarber, R., R. Grishman, P. Tapanainen and S. Huttunen (2000). Automatic acquisition of domain knowledge for information extraction. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*.pp. 940-946