A New Method to Build NLP Knowledge for Improving Term Disambiguation

E. MD. Abdelrahim[1], El-Sayed Atlam[2], R. F. Mansour[3]
[1,3] Computer science department, Faculty of science, Northern Border University, KSA)
[1,2] Dept. of Mathematics, Computer Science Division, Faculty of Science, Tanta, Egypt
[2]Dept. of Information Science and Intelligent Systems, University of Tokushima, Tokushima, Japan 770-8506

***Abstract*—Term sense disambiguation is very essential for different approaches of NLP, including Internet search engines, information retrieval, Data mining, classification etc. However, the old methods using case frames and semantic primitives are not qualify for solving term ambiguities which needs a lot of information with sentences. This new approach introduces a building structure system of natural language knowledge. In this paper all surface case patterns is classified in advance with the consideration of the meaning of noun. Moreover, this paper introduces an efficient data structure using a trie which define the linkage among leaves and multi-attribute relations. By using this linkage multi-attribute relations, we can get a high frequent access among verbs and noun with an automatic generation of hierarchical relationships. In our experiment a large tagged corpus (Pan Treebank) is used to extract data. In our approach around 11,000 verbs and nouns is used for verifying the new method and made a hierarchy group of its noun. Moreover, the achievement of term disambiguating using our trie structure method and linking trie among leaves is 6% higher than old method.**

Keywords—Information Retrieval; NLP Knowledge; Disambiguation; Word Semantics; trie structure

I. INTRODUCTION

Natural language processing (NLP) systems use many dictionaries. In this paper, we discus two types of information. The first is morphological information about morphemes, or words, and their fundamental attributes such as a part of speech [11], and the second is semantic primitive [16][17][28], and so on.

The understanding of implicit in events is of great interest in recent years. Nouns in NL is assumed as real-world entities due to the implicit with nouns in most of work. The lexical nouns name classes of entities, some of which are kinds and some of which are not. This is compatible with the view of compositional semantic in which nouns are viewed as one-place predicate. They are argument-taking functions which take individuals into truth values. On the other hand, verbs are viewed as n-place predicates, functions which take n-tuples into truth values. The (extensional) meaning of any sentence is composed by recursively combining functional terms with quantifiers, operators, and logical connectives.

So first, generic knowledge of events consist of implication is very essential for understanding. For example a person buys something because he wants it. Such knowledge incorporates the implications that buying is enabled by having enough money, and that asking implies that subject of the asking wants something. The second type of knowledge classifies verbs with subject, object, and place into groups which are considered relation. It is important to design an Implicit Inference of nouns and linking group that can efficiently integrate multi-attribute relation.

Implicit inference iinformation is defined by knowing the verbs deepest meaning, determining a deep knowledge about nouns. Also, multi-attribute relation information is defined by a pair of basic words and its record includes the attribute of relation . Consequently, the problems, are a very large space cost for storing all pairs and a high frequent access of pairs and their attribute in the record. A trie, or a digital search structure, must be introduced to the basic scheme since a word is basically a string. Relational information such compound words is formed significantly, and occupies a large spaces in the morphological dictionary. Artificial intelligence (AI) basic knowledge IS-A also depend on term relationships.

A case frame [25][27] is an important technique to solve ambiguity in syntax and semantic analysis [20][21]. Japanese to English, machine translation systems in both direction [25] requires using case frame to build translation dictionaries.

Aoe et al.[1][2][3][4] and Morita et al.[5] introduced a two-trie structure for storing compound words into the compact structure. Morita et al. [6] presented a link trie.

This paper present an implicit inference of nouns, and collect all knowledge about the sentence and make it groups of linking and high frequent access between verbs with subject, object and place. Moreover, by introducing a trie that can define the linkage among leaves, this paper present an efficient data structure. Therefore, the proposed structure defines, multi-attribute relationships between words which can be merged into the same record.

Section II of this paper describes relational information as multi-relations among terms with a case frame of the basic knowledge. The link trie and an integrating morphological is presented in Section III. The proposed method is verified by simulation results in Section IV. In Section V, we discuss conclusions and potential future work.

II. MULTI – RELATIONSHIPS AMONG WORDS

A. Information Of Multi-Attribute Relation

MOR(x) is the morphological information for word x. here we will discuss relational information, call a multi-attribute relation, for a finite of relational attributes briefly.

Multi-attribute relation's information can be defined as a triplet (x, y, Alpha), where x and y are interrelated, and the attribute is Alpha. In natural language processing one can get a variety of attributes, and clearest meaning by using relationships among words as follows.

B. Case frame

To cope with this complexity we have to use the services of some syntactic and semantic information at the same time for the analysis of a sentential structure. The best grammatical framework for this purpose is the case grammar (C. Fillmore in 1968). the semantic primitive shown in Table 2 is utilized to determine which kind of noun can be in which case slot. For instance, the verb eat load a noun connected with one of the semantic primitive animal as the cause of the verb, and noun of semantic code eatable stuff as an object. This case slot determination is specified for each handling of all verbs in a dictionary.

The information to be inserted in the dictionary record differs depending on each part of speech, but in general include this kind of information: head word, number of character of words end, alternate, root word, correlated words, morphological piece of spoken language, conjugation, prefix information, area code, grammatical part of speech, sub-categorization of piece of speech, patterns case, feature, model, option, semantic primitives, co-occurrence information (adverb, predicative modifier), idiomatic expressions, degrees, degrees of nominality and so on.

Here, Verbs and nouns case pattern is one of the important information. We have renowned over 30 instances, see Table 1. Each case slot in a pattern of verb use include semantic information about the noun, which could be seen in the slot. The noun has the matching semantic code in an entry. We have renowned over 50 semantic primitives (codes) in Table 2.

TABLE I. CASE RELATIONS USED IN THE ENGLISH DEPENDENCY STRUCTURE [M.NAGAO,ET.AL[13]]

(1) Subject	(17) Attribute
(2) Object	(18) Cause
(3) Recipient	(19) Tool
(4) Origin	(20) Material
(5) Partner	(21) Component
(6) Opponent	(22) Manner
(7) Time	(23) Condition
(8) Time-From	(24) Purpose
(9) Time-to	(25) Role
(10) Duration	(26) Content
(11) Space	(27) Range
(12) Space-From	(28) Topic
(13) Space-To	(29) Viewpoint
(14) Space-Through	(30) Comparison
(15) Source	(31) Accompany
(16) Goal	(32) Degree
	(33) Predicative

TABLE II. SYSTEM OF SEMANTIC PRIMITIVES FOR NOUNS (NAGAO ET AL., 1986)

NATION & ORGNAZATION			
ANIMATE	1-HUMAN.PROFFSION	PHENOMENON	1-TURAL PHENOMENON
	2-ANIMAL		2-PHYSCAL PHENOMENON
	3-PLANT		3-POWER&ENERGY
	4-OTHERS		4-PHYSIOLOGICAL PHENOMENON
			5-SOCIAL PHENOMENON
INANIMATE	1-NATURAL SUBSTANCE		6-SOCIAL SYSTEM
	2-PARTS MATERIALS		7-OTHERS
	3-ARTIFICAL PRODUCT		
	4- SYSTEM	FEELING	1-FEELING MENTAL
	5-OTHERS		2-THINKING
			3-OTHERS
ABSTRACT	1-INTERLLECTUAL PRODUCT		
PRODUCT	2-INTERLLECTUAL TOOL	ACTION	1-DOING
			2-MOVING
	3-INTERLLECTUAL MATERIALS		
			3-OTHERS
	4-INTERLLECTUAL GOODS		
	5-OTHERS	MESURMENT	1-NUMERIC
			2-MEASURABLE PROPERTY
PART	1-PARTS ELEMNET		3-STANDARD
			4-UNIT
	2-ORGANS OF HUMAN OR ANIMAL		
			5-OTHERS
	3-OTHERS		
		PLACE LOCATION	
ATTRIBUTE	1-NAME OF ATTRIBUTE		
	2-RELATION	TIME	1-TIME POINT
	3-SHAPE		2-TIME DURATION
	4-STATE		3-TIME PROPERTY
	5-PROPERTY		4-OTHERS
	6-OTHERS		
		OTHERS	

In view of the Machine Translation (MT) example by [M. Nagao, et. al. [26],[27] as in Table 1, the semantic primitive is employed to determine which kind of noun can be in which case slot. For instance, the verb eat requires a noun linked with one of the semantic primitive animal as the cause of the verb, and noun of semantic code eatable substance as an object. That case slot determination is specified for each use of all verbs in dictionary.

The VERB and NOUN (OBJECT, PLACE) relation relations are defined as follows:

Ahmed reside in EGYPT	*<VERB –PLACE>*
Ahmed speak Arabic	*<SUB. – VERB>*
Cat eat food	*<VERB – OBJ.>*
Fish live in water	*<VERB – OBJ.>*
Ibrahim treat sickness	*<SUB. – VERB>*

In the following sub section, more detailed study can be carried out with examples to have an implicit meaning of term disambiguation.

Example [1]. SEMANTIC(“Chocolate”) [PLACE \ MOUNTAIN]

Sentence: Jhon will climb the Chocolate in the next winter holiday.

(ACTOR: Jhon, HUMAN)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “climb”, then the noun “Chocolate” is a PLACE where HUMAN will climb on it. Therefore, SEMANTIC(“Chocolate”) in the previous sentence is [PLACE \ MOUNTAI][9-15].

Example [2]. SEMANTIC(“Chocolate”) [FOOD \ EAT]

Sentence: Hala eats Chocolate.

(ACTOR: Hala, HUMAN)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “eats”, then the noun “Chocolate” is an EATABLE MATERIAL that HUMAN will eat. Therefore, SEMANTIC(“Chocolate”) in this previous sentence is [FOOD \ EAT].

Example [3]. SEMANTIC(“Chocolate”) [PRODUCT MOBILE PHONE]

Sentence: Data is organized in Chocolate.

(ACTOR: Data, INTELGENT PRODUCTs)
(OBJECT: Chocolate)

As in Table 2 of semantic primitives, we will find that “Chocolate” is an OBJECT, and by the information of verb “organized”, then the noun “Chocolate” is an INTELGENT PRODUCTs that can be organized data on it. Therefore, SEMANTIC(“Chocolate”) in this sentence is [PRODUCT \ MOBILE PHONE].

Examples show in sentence ambiguities, WSD can be carried out based on the clear semantic primitives in sentences. However, in the case of context ambiguities, although the sentence includes semantic primitives, context ambiguities are still hard to be solved, Appendix A.

C. Implicit Inference of a Noun

It is very essential to have systematic study on the verbs and nouns, to have a deep knowledge. Due to implicitly in the events, a generic knowledge is necessary. By creating a verb semantic representation in the case frame, we can get more information about noun. A detailed study has been carried out with many examples to get nouns implicit inference as follows:

Example [1]: Mr. Atlam eat fried fish in a restaurant.

For a case frame of this sentence;

(ACTOR: Mr. Atlam)
(OBJECT: fried fish)
(LOCATION: Restaurant)

We notice that, a noun has just semantic primitive. For example in this example, we find that fried fish is one kind of food. This means that by using Table 2 (semantic primitive) that food one PLANTS and plants have no knowledge about ‘eatable material.’ Also, a restaurant by Table 2 just a LOCATION, and has no knowledge about ‘eating place.’

By using implicit inference (deep information of a verb), we find a SEMANTIC_REFER of slot, which indicates that the frame may possibly refer to the semantic depiction of that slot. The knowledge of the verb eat in this example refers to knowledge of fried fish and a restaurant, then an object fried fish is referred to ‘eatable material’ and a restaurant is referred to ‘eating place.’

Example 2: Mr. Samouda swims in a river.

For another case frame of the sentence;

(ACTOR: Mr. Samouda)
(LOCATION: River)

By using the semantic primitive of noun Table 2, we find in this example that a river refers to only LOCATION, and has no knowledge about ‘swimming place,’ and Mr. Samouda refers to the HUMAN.

But by employing the deep information of the verb we see a slot of SEMANTIC_REFER indicating that the frame may refer to the semantic description of that slot. The knowledge of the verb swim in this example is refers to the knowledge of a river, then the place a river is referred to ‘swimmable place’ where the HUMAN swim, and if there is relation that LOCATION has OBJECT, then the dynamics knowledge that a river has water.

Example 3]. Mr. Atlam wants to buy a computer from a store.

For this case frame of this sentence;

(ACTOR: Mr. Atlam)
(OBJECT: Computer)
(LOCATION: Store)
(TOOL: $10,000)

By using the semantic primitive for nouns Table 2, this refers to a computer which is an ARTIFICIAL PRODUCT, a store refers to the place location, and $10,000 just TOOL (has no information about the price of computer).

But by employing implicit inference, we find the slot of SEMANTIC_REFER indicating that the frame may possibly refer to the semantic depiction of that slot. The knowledge of the verb buy in this example is refers to knowledge of a computer, a store, and $10,000, then the knowledge refers that Mr. Atlam is enabled, so by having money and the cost is $10,000, that Mr. Atlam intends to use what he buys, also store is referred to 'buyable place', if there is a relation that LOCATION has OBJECT, then a store has a computer.

By using verb information in the case structure, implicit knowledge of nouns can be derived. By extending this knowledge, we can build some linkage groups between a subject with a verb, a verb with an object, and a verb with a place. We can write the same typical sentence, as follows:

1) My wife eat some meat in a restaurant.
2) My father eat some rice in a restaurant.
3) Mr. Mohammad swims in the sea.
4) My son swims in a pool.
5) My daughter buys a toy from the store.
6) Mr. Mathew buys a television from the store.
7) My Mother buys some fruits from the market.
8) The students drink a glass of juice in his house.
9) Math teacher drinks some coffee in the school, and so *on.*

By collecting a large number of this examples, we could build the following groups: the linking between nouns and verbs is shown in the figure 1, and this group is arranged from down to up depending on the strong and has the weak relation. This means that the relation between nouns and high-leaky verbs, such as talk, think, speak, and so on. They are verbs of a higher animal action, and strong verbs. But another is general verbs, such as eat, drink, and so on, or a weak verb, as follows:

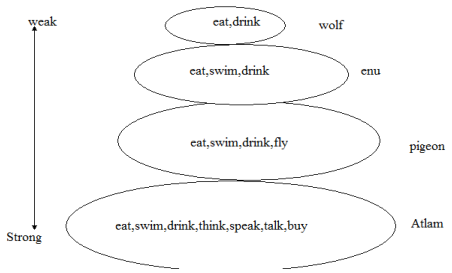

Fig. 1. Group of Link between a Subject and Verbs

Although each knowledge dictionary in primitive systems is built separately, almost all modern natural language applications become more complicated combining the above relationships. For this reason, it become necessary to design a fast and compact structure to be efficiently integrated with any of multi-attribute relation.

Since information about multi-attribute relation is defined by a pair of basic words and its record as well as the attribute of relation, the problems become a very large space cost for storing all pairs and a high frequent access of pairs and their attributes in the record. Since a word is basically a string, a trie, must be added to the basic scheme representation

D. Compound Word

The triple <x, y, α> is called Compound Word relations which indicates that x composite with y to give new information. By using case frame relation <tool> + <Verb> → of computer processing, and <Subject> + <Verb> → of language processing are called compound word relation. By using this case frame relation the clearest meaning and information about word can be extract rather than the single one, another example as follow:

Information Retrieval

Natural language.

III. LINK TRIE (LT) FUNCTION

A. Tries and Efficient Representation of Verb and Noun Linkage

Trie is an n-array tree [2], [10], [11], [15] having n-place vectors as nodes with components corresponding to digits or characters. For confusion avoidance between keys like the and then, let us insert a special end marker; # to the end of all keys, so no prefix of a key can be a key itself [1]. Let K be a keys set. Each path in the trie starting from the initial node (root) to a leaf corresponds to a key in K. Therefore, the nodes of the trie correspond the prefixes of keys in K. A trie definition is as follows [3], [4], [5].

1) S is a limited set of nodes, represented as a positive integer.
2) I is a limited set of input characters, or symbols.
3) g is a goto function from S □ I to S ∪{fail}.

This means that, a node r is in F if and only if there is a path from 1 to r reads some string x in K. A move titled with a (in I) from r to t means g(r, a) = t. The nonexistence of a move means stoppage (failure). Figure 2 shows a trie example for eleven words with '#', where enclosed in a square nodes will be later discussed. The key 'Atlam#' retrieving can be done by applying the transitions g(1, 'a') = 3, g(3, 't') = 22, g(22, 'l') = 14 , g(14, 'a') = 32, g(32,'m')=15, and g(15,'#')=2, sequentially.

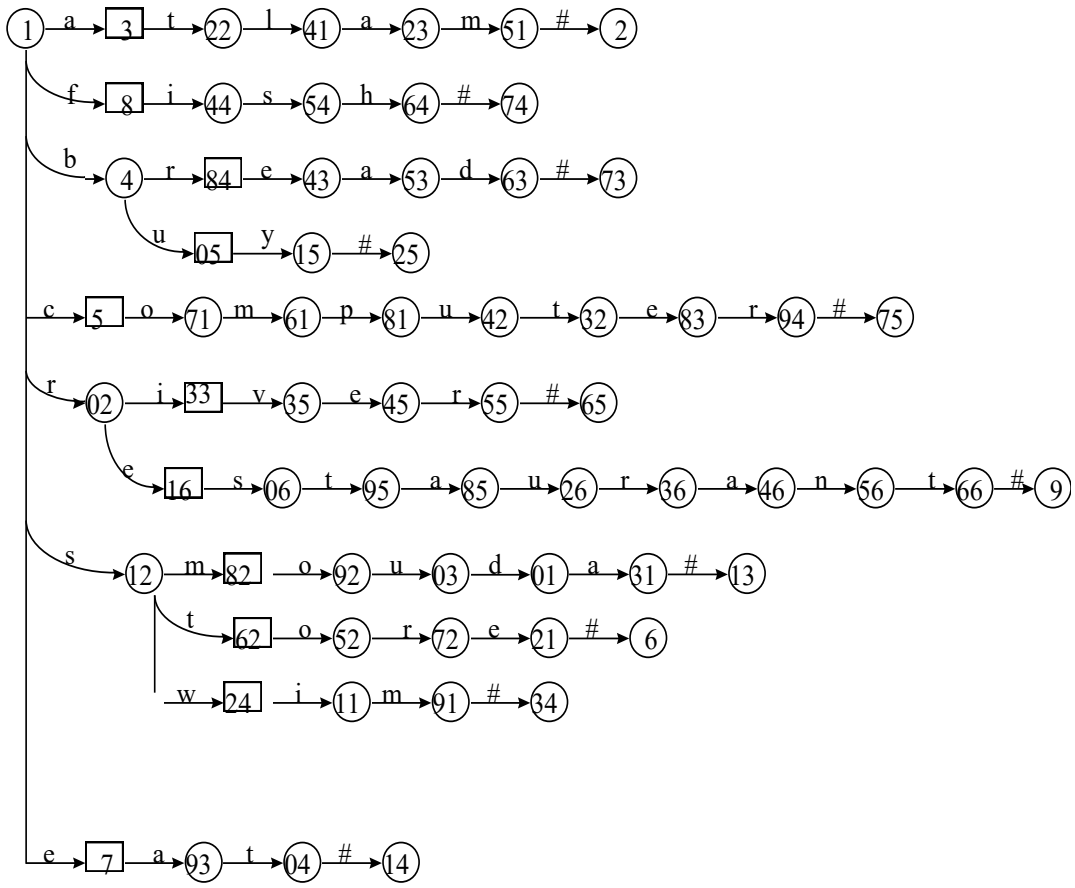

Fig. 2. Example of Trie Structure

B. *Link Trie(LT) Function [K. Morita, 6]*

Term Relationships Definition

Assume (X, Y, R) is the relation R between termss X and Y. With tries, there is one-to-one correspondence between leaves and keys, so we can define its link trie by linking leaf s for X and leaf t for Y. In such case, the definition of function LINK is t ∈ LINK(s) and the relation by the record R ∈ CONTENTS(s, t). Link trie is the trie including the function LINK and CONTENTS. Link information for figure 2 is shown in Table 3.

We can see the relationship between Atlam as a subject and buy as a verb by the trie and there exist one-to-one correspondence, the leaf 2 correspondence key Atlam and leaf 52 correspondence key buy, and link function is defined by 52 ∈ LINK(2) and the record (<subject>, <verb>) ∈ CONTENTS(2, 52). We can see the relationship between words (<verb>, <object>) and (<verb>, <place>), as follows:

Retrieval Algorithm

For the relationship (X, Y, R), the proposed retrieval algorithm (i): retrieve Y and R from X, (ii): retrieve R from X and Y.

For LT and for key X, the function GET_LEAF(LT, X) gives the leaf for X# and gives fail if LT has no X#. The function GET_LEAF (LT, "store") gives leaf 6 in Figure 2.

For the relationship (X, Y, R), the following ALGORITHM returns leaves s for X# and t for Y# if they are recorded in the trie. s and t could be processed to recover CONTENTS(s, t) including relationship R. If any of s or t is not recorded in the trie, then ALGORITHM outputs s = t = 0.

[ALGORITHM]

start
s← GET_LEAF (LT, X);
t← GET_LEAF (LT, Y);
if (s = fail or t = fail) then output s = t = 0;
if ((t ∈ LINK(s) and R ∈ CONTENTS(s, t)) then output s and t;
end;
(Algorithm End)

C. *System Frame work*

Figure 3, shows the frame work of our approach by Searching for Some English Textbook &Papers, concerning with Cross Language Information, Classification Summarization, and Noun Extraction from the Penn

Treebank• Extract compound noun after stemming and use stop word dictionary , from large Corpus. Moreover, Extract the linkage between verb with noun, verb with place , and verb with place , by using part of speech dictionary, and make linkage group and high leaky relation between them. By using this frequent and high leaky relation we can make disambiguate for word, where the surrounding words frequently associated with a sense are used to disambiguate a word.

TABLE III. EXAMPLES OF INFORMATION LINK

X	S	LINKS	CONTENTS(s, t)
Atlam	2	{41,52}	CONTENTS(2,41)={<subject>,<verb>}
			CONTENTS(2,52)={<subject>,<verb>}
Smouda	31	{2,43}	CONTENTS(31,2)={<subject>,<subject>}
			CONTENTS(31,43)={<subject>,<verb>}
eat	41	{37,47,9}	CONTENTS(41,37)={<verb>,<object>}
			CONTENTS(41,47)={<verb>,<object>}
			CONTENTS(41,9)={<verb>,<object>}
buy	52	{57,6}	CONTENTS(52,6)={<verb>,<place>}
			CONTENTS(52,57)={<verb>,<object>}

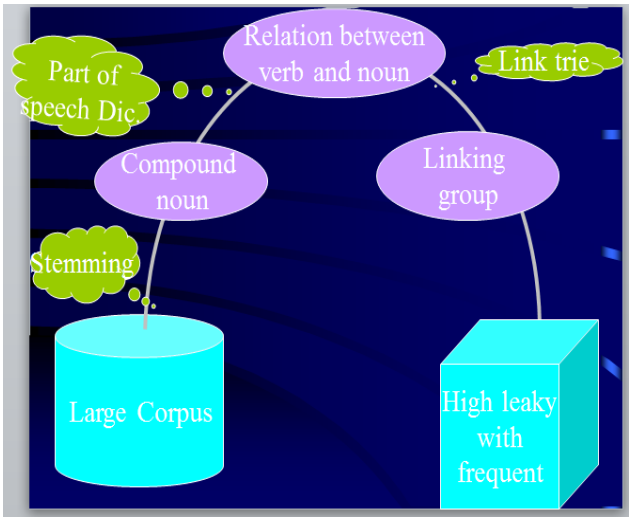

Fig. 3. System Framework

D. Semantic Field Information

As Section 2.1 discussed that some words have many semantic meaning. Therefore, various semantic(x) usually appears in various branches. Table 4 shows that [PLACE \ MOUNTAIN] is in fields <TRIP \ AMERICA>, and <SPORTS \ MOUNTAIN CLIMBING>. [FOOD \ EAT] is in <FOOD \ SUPERMARKET>. [PRODUCT \ MOBILE PHONE] is in <COMPANY \ TELPHONE SHOPE>. Therefore, words with various fields in the context could be utilized to discriminate the semantic(x).

TABLE IV. EXAMPLES OF RELATIONSHIPS BETWEEN SEMANTICS AND FIELDS

SEMANTIC("Chocolate")	semantics Ambiguities	Field
[PLACE \ MOUNTAIN]	After dinner, our manager eat some snacks. Both Hala and Jhon usually eat Chocolate, because they use Chocolate.	<TRIP \ AMERICA>, <SPORTS \ MOUNTAIN CLIMBING>,
[FOOD \ EAT]		<FOOD \ SUPERMARKET>
[PRODUCT \ MOBILE PHONE]		< COMPANY \ TELPHONE SHOPE >

IV. AUTOMATIC KNOWLEDGE GENERATION FOR AN UNKNOWN WORD

This section describe how to get more information & new knowledge from case-frame storing by using trie structure and linking between leaves, perhaps by keeping links between them to reflect some relationships. e.g. Jhon * is unknown word

Context (case frame)

Level1: Jhon * eats apple, Jhon IS – A animal?, Jhon is similar to dog, or human

Level2: Jhon * buys computer, Jhon IS – A human.

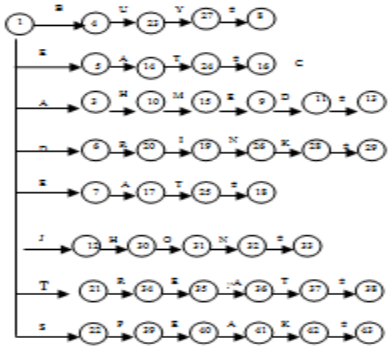

Fig. 4. Trie structure

By using this information as common knowledge, we will show later by using trie structure and link trie, how we can know new automatic & variety relation from this common knowledge. This new knowledge are very useful in NLP, because it make a text more readable and understandable for human, this new knowledge can be combined to provide additional useful <IS –A> hierarchical information, as follow:

In this examples with <SUB. – VERB> relation using the information:

1- Ahmed treat the illness *2- Ahmed cure the sick*
2- Ahmed eat food *4- Ahmed speak with the nurse*
3- Ahmed drink milk *6- Jhon eat orange*
4- Jhon drink tea *8- Jhon speak with his teacher*
5- Cat eat food *10- Cat drink water*

and create the structure of trie as Figure 4. and links between leafs in this trie, we can build the linking as in table 5 from the given information as follows:

TABLE V. TRIE LINK FOR <SUB. –VERB> RELATIONSHIP

X	s	RELATION(s)	ATTRIBUTE(s, t)
Jhon	33	{2,18,8,29}	ATTRIBUTE(33,2)= <SUB. -- VERB > ATTRIBUTE(33,18)= <SUB. -- VERB> ATTRIBUTE(33,8)= <SUB. -- VERB> ATTRIBUTE(33,29)= <SUB. -- VERB>
Ahmed	13	{2,8,29,18,38}	ATTRIBUTE(13,2)= <SUB. -- VERB> ATTRIBUTE(13,8)= <SUB. -- VERB> ATTRIBUTE(13,18)= <SUB. -- VERB> ATTRIBUTE(13,38)= <SUB. -- VERB> ATTRIBUTE(13,29) = <SUB. -- VERB>
Cat	16	{2,29}	ATTRIBUTE(16,2) =<SUB. -- VERB> ATTRIBUTE(16,29) = <SUB. -- VERB>

Next using this automated linking information, one can understand from this linkage that things which can eat and drink only and cannot speak and buy (i.e. eatable & drinkable only) is Animals, also things which can eat ,buy, drink, and speak and cannot treat sickness (i.e. buyable, speakable , eatable, drinkable only) is a provoke (normal) human, and the man who can eat food , drink drinks , buy goods , speak languages and have the ability to care for sickness (treatable) is a doctor. And we can create also this group as in Figure 5.

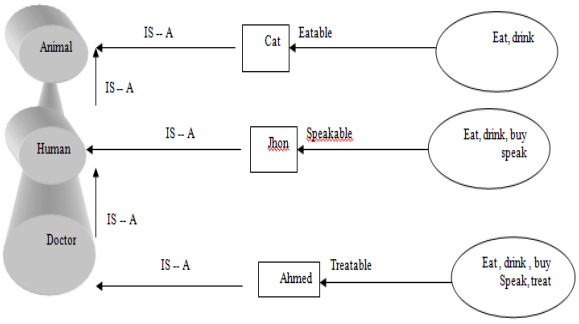

Fig. 5. Hierarchy and clear knowledge extract from link trie

From this link trie, we can get the < IS -- A> hierarchy relationship that *Doctor is a human, and human is an animal.*

V. SIMULATION RESULTS

A. Experimental data and information

99,714 statements from tagged corpus (Pan TreeBank), having diverse of features, is implicated in this experiment.

Data Set 1:

About 11,970 subject-verb case relationship and about 2,514 of verb-object relationship, and 679 verb-places are used. Due to high frequent access of pairs, we could not take them up. See Table 6.

TABLE VI. HIGH FREQUENT ACCESS OF PAIRS & LINKING

Verb	*Subject Frequent*	*Object Frequent*	*Place Frequent*
buy	holders (41) manufacturers (41) consumers (19) worldbank (13) people (12) company(10)	food (28) computer (27) recorder (18) clothes (14)	company(24) institutions (21) market (14) farm (10)
change	Jhon (80) farmers (35) rules (25) money (30)	money (70) shares (30) rate (43) dollar (50)	bank (31) hotel (22) company (13)
sell	traders (24) tourists (19) farmars (15) jhon (13) foreign (10)	car (61) currencies (41) computer (31) bus (50)	institutions (51) company (23)
read	Jhon (80) bank (30) company (48)	money (70) program (30) role (20) shirt (7)	bank (31) company (45) office (34)
eat	Jhon (100) people (98) farmars (70) tourists (19)	apple(150) orange(140) Duck (50)	office (34) resturant (70) company(12)
draw	foreign (100) farmars(67) jhon(50)	book(100) shirt(50) duck (40)	bank (31) institutions(20)

Data Set 2:

Utilizing case frame with trie structure to present a lot of relationships between words as shown in section II.

Data Set 3:

Employing trie structure with linking trie among leaves for additional information as shown in Figure 4, and Table 3.

Data Set4:

Restrict 10 group of typical verbs and objects from Data 1, as in Table 5.

Result [1].

By employing Data 1,2,3. We can establish an automated generation of hierarchy of relations among words as shown in Figure 5 and Algorithm 2.

Result [2].

By gathering this data and establishing relationships among verbs and other kinds of keys with link trie, we can see the hierarchy group. Figure 6 for example shows the subject and verb linkage.

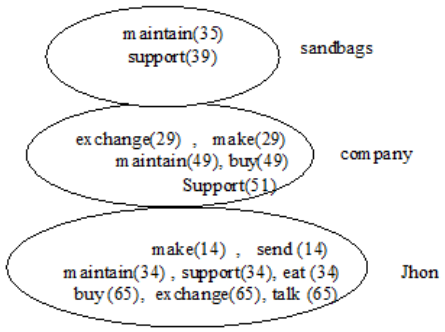

Fig. 6. Subject and verb Linkage group

Utilizing high-leaky this indicates that ***sandbags*** are maintainable and supportable but not exchangeable or buyable, ***company*** exchangeable , buyable supportable, maintainable , but not eatable ,talkable, but ***Jhon*** can , maintain, support, , buy, exchange, eat, talk.

Result [3]. Disambiguation

Figure 7 shows how the algorithm fare with sentences containing ambiguous element, be able to handle many such cases, as will be illustrated here. Consider the pair of sentences below:

1) Investment company support the bank.
2) The sandbags support the bank.

By this three sentence we show the semantic meaning of the word bank have two meaning financial house & edge of rive and by use more information about the word bank by another verbs, we can change the case to disambiguation case. As follow: The first approach: semantic meaning of bank is financial house in the first sentence, this by using another verbs to declare this meaning as in these sentence say : Jhon exchange from bank. and for more information about bank we can say that : Bank buy money . By this more information we find all sentence speak about money this implies more disambiguate for word bank and now the clear semantic is financial institution .the second approach: semantic meaning of bank is edge of river in the second sentence, this by using another verb to declare this meaning as in these sentence: Sandbags maintain bank, and for more information about bank we can say: Bank maintain river i.e. By this more information we find all sentence speak about hold up physically this implies more disambiguate for word bank and now the clear semantic is edge of river.

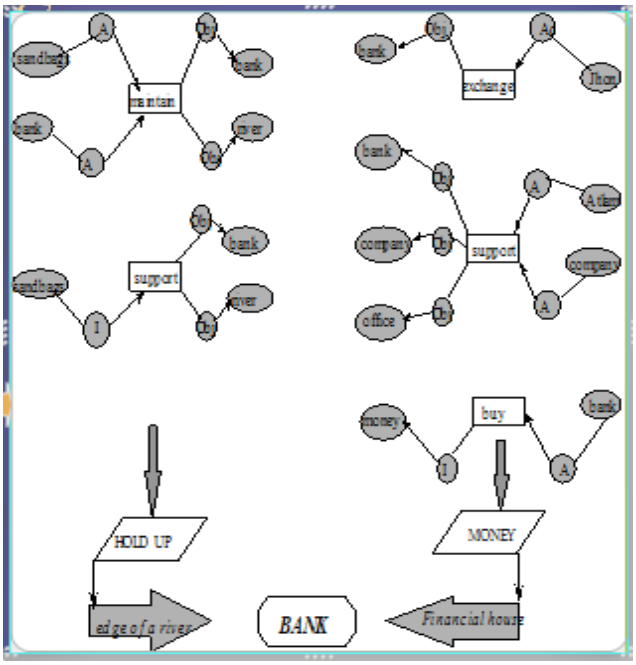

Fig. 7. Example of disambiguation

Result [4]. The accuracy of experimental results is defined as:

$$\text{Accuracy} = \alpha/\beta$$

Where α is the number of words disambiguated correctly and β is total number of ambiguous words.

Table 7 summarizes the experimental observation of using the old method (TM) and new method using the trie structure and linking trie between leaves. All English terms of our experimental are in Appendix A.

In Table 7, both performed with accuracies 73% for TM and 79% by new method using the link trie between leaves which is significant than the value that can be obtained by TM one. This means that, our new approach is viable in solving term ambiguities.

TABLE VII. EXPERIMENTAL RESULTS OF THE WHOLE WORDS

		TM*	NM*
		EN*	EN
Number of terms used in the Experiments			38 372
Sentence Ambiguities	No. of Ambiguous Words	520	520
	No. of Unambiguous Words	380	410
	Accuracy	73%	79%

* EN = English TM*= Traditional method NM*= New method (Using Linkage)

VI. CONCLUSION

In this paper, we proposed a new approach for building structure system of natural language knowledge. In this paper all surface case patterns are classified in advance with the consideration of the meaning of noun. Moreover, this paper introduces an efficient data structure using a trie which define the linkage among leaves and multi-attribute relations. By using this linkage multi-attribute relations, we can get a high frequent access among verbs and noun with an automated generation of hierarchical relationships. In our experiment a large tagged corpus (Pan Treebank) is used

to extract data. In our approach around 11,000 verbs and nouns is used for verifying the new method and made a hierarchy group of its noun. Moreover, the achievement of term disambiguating using our trie structure method and linking trie among leaves is 6% higher than old method. The preliminary result of our method shows a good promise, because the extracted information structures of a special database, can be extended by a more large input of data and more general relations from a large information corpus. The results of disambiguating the word ambiguities are much better than that of case frames. Experimental results also show that enough distinctive terms can help determine the semantic sense of a word in a specific context. The preliminary syntactic analysis can be achieved by many natural language processing system, we will be able to obtain more precise semantic information from the syntactic resource. Moreover, the accuracy of disambiguating words by our method using trie structure and linking trie between leaves is 6% higher than traditional method. Future work could focus in using context analysis to improve disambiguates of words. Extract Arabic keyword By using stop word Dictionary and stemming rule, from large Arabic Corpus with Classification for Arabic text by using Classification engine.

ACKNOWLEDGMENTS

The authors wish to acknowledge the approval and the support of this research study by the grant no. 5-9-1436-5 from the Deanship of Scientific Research in Northern Border University, Arar, KSA.

REFERENCES

[1] A.V. Aho, J. E. Hopcroft, and J. D. Ullman, “Date Structure and Algorithm,” Addison-Wesley, Reading, Mass., pp. 163-169, 1983

[2] M. Ai-Suwaiyel and E. Horowitz, “Algorith for Trie Compaction,” ACM Trans. IEICE, Vol. J76, D-II, No. 11, pp. 243-263, 1984

[3] J. Aoe, “An Efficient Digital Search Algorithm by Using a Double-array Structure”, IEEE Trans. Software Eng., Vol. 15, No. 9, pp. 1066-1077, 1989

[4] J. Aoe, K. Morimoto and T. Sato, “An Efficient Implementation of Trie Structure,” Software-Pract. & Expr. Vol. 22, No. 9, pp. 695-721, 1992

[5] J. Aoe, K. Morimoto, M. Shishibori, and K. Park, “A

[6] Trie Compaction Algorithm for Larg Set Keys”, IEEE. Trans. on Knowledge and Date Eng., Vol. 8, No. 3, 1996

[7] J. Aoe, K. Morita, H. Mochizuki, and Y. Yamakawa, “An Efficient Retrieval Algorithm of Collocational Information Using Trie Structures" (in Japanese), Transactions of the IPSJ, Vol. 39, No. 9, pp. 2563-2571, 1998

[8] J. Aoe, String Pattern Matching strategies, 1994.

[9] E. Brill, “A Case Study in A Part of Speech Tagging”, Computational Linguistics, Vol.21, No. 4, pp. 1-37, 1995.

[10] Abdunabi Ubul, El-Sayed Atlam, Hiroya Kitagawa, Masao Fuketa Kazuhiro Morita and Jun-ichi, Aoe An Efficient Method of Summarizing Documents Using Impression Measurements, COMPUTING AND INFORMATICS Journal, Volume 32, No. 2, 2013. Atlam E.-S, , Morita, K.., Fuketa M, Aoe, & J.. A new for selecting English compound terms and its knowledge representation. Information Processing & Management Journal, 38(6), 807–821. (2002

[11] Atlam, E.-S., Fuketa, M., Morita, K., & Aoe, J. Document similarity measurement using field association terms. Information Processing & Managemen Journal, 39(6), 809–824. (2003).

[12] Atlam, E.-S, Elmarhomy, G., U. M. Sharif, Fuketa, M., Morita, K., & Aoe, J. Improvement of building field association term dictionary using passage retrieval. Information Processing & Management Journal, 43, 1793–1807. (2007).

[13] M. E. Abd El-Monsef, El-Sayed Atlam and O. El-Barbary, Combining FA Words with Vector Space Models for Arabic Text Categorization, An International Journal of INFORMATION, Vol. 6, No.(6A), pp.3517-3528, 2013.

[14] Atlam El-S. and El-Barbary O., Arabic Document Summarization using FA Fuzzy Ontology, , International Journal of Innovative Computing, Information and Control, 2014.

[15] Atlam El-S., Improving the Quality of FA Word Dictionary based on Co-occurrence Word Information and its Hierarchically Classification, , International Journal of INFORMATION Vol.17, No.2, February, 2014.

[16] K. Dahlgren, Naive semantics for Natural Language Understanding, 1982.

[17] W. B. Frakes, Information Retrieval Data Structure &Algorithms, 1992.

[18] E. Fredkin, “Trie Memory”, Commun. ACM., Vol. 9, No. 2, pp. 490-500, 1960

[19] D. E. Knuth, “The Art of Computer Programming”, Vol. 3, Sorting and Search, pp. 481-505, 1973

[20] F. Fukumoto, “Disambiguating preposition phrase attachment using statistical information” , NLPRS., Vol. 34, No. 2, pp. 752-757, 1995.

[21] Y. Jin, and Y. Tackkim, “Noun-sense Disambiguation from the Concept Base in MT” , NLPRS., Vol. 32, No. 2, pp. 357-362, 1995.

[22] J. Kupiec “A Robust Linguistic Approach For Question Answering Using An On-Line Encyclopedia, In proceedings of 16th ACM SIGIR international conference, pp. 181-190, 1993.

[23] H. Li., and N. Abe, “ Clustering Words With the MDL Principle”, Journal of Natural Language Processing, Vol. 4, No. 2, pp. 71-88, April 1997.

[24] K. Lim and M. Song, “Morphological Analysis with Adjacency Attributes and Word Dictionary”, In proceedings of the international conference on computer processing of oriental language, pp. 263-268, 1994.

[25] D. W. Loveland, Natural Language Parsing system, 1987

[26] M. Nagao, J. Tsujii, and J. Nakamura, “Machine Translation from Japanese to English”, Vol. 74, No. 7, pp. 993-1012, 1986

[27] A. Oishi, and Y. Matsumoto, “ A Method for Deep Case Acquisition Based on surface Case Pattern Analysis”, NLPRS., Vol. 34, No.2, pp. 678-684, 1995.

[28] T. A. Standish, Data Structure Techniques, 1981

[29] R. E. Tarjan and A. C. Yao, ”Sorting a Sparse Table”, Commun. ACM., Vol. 22, No. 11, pp. 606-611, 1979.

[30] T.Takenobu and I.Makoto, “Text Categorization Based on Weight Inverse Document Frequency”, SIG-IPSJ, pp. 33-39, 1994.

[31] T. Satomi, Atlam El-S., Morita K., Fuketa M. and Jun-ichi Aoe Context Analysis Scheme of Detecting Personal and Confidential Information, , International Journal of Innovative Computing, Information and Control, Vol.8, 5(A), pp.3115-3134, 2012

[32] A. Utsumi, K. Hori, and S. Ohsuga “ An Affective- Similarity-Based Method for Comprehending Attributional Metaphors” , Journal of Natural Language Processing, Vol. 5, No. 3, pp. 3-30 July 1998.

APPENDIX A: ENGLISH WORDS IN EXPERIMENTS

Number	English words	Semantics
1	Kilimanjaro	[Place \ Mountain] [Food \ Drink] [Product \ Software]
2	puma	[Cougar] [Car] [Sportswear]
3	beetle	[Animal \ Insect] [Car]
4	polo	[Sports] [Clothing]
5	hawk	[Animal] [Aircraft in military] [Sports club]
6	queen	[Person] [Animal] [Games]
7	jaguar	[Animal] [Car]
8	apple	[Fruit] [IT company]
9	panda	[Animal] [Chinese car] [Fast-Food] [Software]
10	fish	[Animal] [Food]
11	blackberry	[Fruit] [Mobile phone]
12	Barcelona	[Place \ City] [Sports club]
13	office	[Working place] [Software]
14	thunderbird	[Animal] [Software] [Food \ Drink] [Aircraft in military]
15	tree	[Plant] [Structure in computer science]
16	tiger	[Animal] [A name of golf player] [Car] [Beer]
17	dove	[Animal] [Aircraft] [Food \ Chocolate] [Toiletry]
18	Mont Blanc	[Place \ Mountain] [Writing instruments and accessories]
19	chocolate	[Food] [Place \ Mountain] [Mobile phone]
20	window	[Object] [Software]
21	match	[Tool \ Fire] [Game] [Japanese car]
22	branch	[Plant] [Structure in computer science]
23	Liverpool	[Place \ City] [Sports club]
24	phoenix	[Film] [Sports club] [Television , Broadcasters]
25	Oracle	[Ancient text] [Software company]
26	cobra	[Snakes] [Aircraft] [IT products]
27	rocket	[Vehicle, missile, aircraft] [Sports club]
28	maverick	[Animal] [Sports club] [Car]
29	mustang	[Animal] [Aircraft in military] [Car]
30	QQ	[Messaging program] [Chinese car]
31	lotus	[Plant] [Car]
32	Amazon	[Geography, river] [IT company]
33	crocodile	[Animal] [Aircraft in military]
34	penguin	[Animal] [Clothing] [Sports club]
35	virus	[Program] [Infectious agent]
36	bridge	[Architecture] [Sports game] [Hardware]
37	line	[Object \ Product] [Formation] [Calling] [Cord]
38	bass	[Musical sense] [Animal \ Fish]
39	cone	[Part of tree] [Sharp of object] [Part of eye]
40	interest	[Curiosity, attraction] [Advantage] [Financial] [Share]
41	taste	[Preference] [Flavor]
42	sentence	[Punishment] [Set of words]
43	train	[Object \ Series] [Movement \ Prepare]
44	book	[Object \ Published document] [Movement]
45	bank	[Institution] [Architecture \ Ground]
46	serve	[Movement \ Ball game] [Movement \ Food]
47	dish	[Food \ Meal] [Receptacle]