Collocation Extraction Performance Ratings Using Fuzzy Logic

Momtaz Thingujam *chur55@yahoo.co.in*
Department of Computer Science
Manipur University
Manipur, 795001, India

Ak.Ashakumar Singh *ashakumars8@gmail.com*
Department of Computer Science
Thoubal College-Manipur University
Manipur, 795138, India

Abstract

The performance of Collocation extraction cannot quantified or properly express by a single dimension. It is very imprecise to interpret collocation extraction metrics without knowing what application (users) are involved. Most of the existing collocation extraction techniques are of Berry-Roughe, Church and Hanks, Kita, Shimohata, Blaheta and Johnson, and Pearce. The extraction techniques need to be frequently updated based on feedbacks from implementation of previous methods. These feedbacks are always stated in the form of ordinal ratings, e.g. "high speed", "average performance", "good condition". Different people can describe different values to these ordinal ratings without a clear-cut reason or scientific basis. There is need for a way or means to transform vague ordinal ratings to more appreciable and precise numerical estimates. The paper transforms the ordinal performance ratings of some Collocation extraction techniques to numerical ratings using Fuzzy logic.

Keywords: Fuzzy Set Theory, Collocation Extraction, Transformation, Performance Techniques, Criteria.

1. INTRODUCTION

There is no widely accepted definition of collocation. More discussions are going on in the linguistics literature on the exact nature of collocation [1]. It is necessary to ensure generation of lexically correct expressions. Collocations are abundant in language and vary significantly in terms of length, syntactic patterns and offset [2]. Measurement ratings of these techniques are ordinal and are subject to ambiguity. This means that these ratings have some elements of uncertainty, ambiguity or fuzziness.

When humans are the basis for an analysis, there must be a way to assign some rational value to intuitive assessments of individual elements of a fuzzy set. There is need to translate from human fuzziness to numbers that can be used by a computer.
Some researchers in natural language processing have proposed computationally tractable definitions of collocation accompanying empirical experiments seeking to validate their formulation such as [3-11] recently.

Berry-Roughe (1973) uses the expected frequency of two words mentioning the slight modification in the window size of the word [3]. **Church and Hanks** (1990) measures the co-occurrence of two words and it becomes unstable when the counts of the words are small [4]. **Kita et al.** (1994) used the idea of the cognitive cost of processing a sequence of words [5]. The technique of **Shimohata et al**. (1997) is capable of extracting both interrupted and uninterrupted collocations [6]. **Blaheta and Johnson** (2001) technique has the effect of trading recall for precision of the words [10]. **Pearce** (2001) technique is a supervised technique based on semantically compositional words [11].

Lofti A Zadeh introduced **Fuzzy Set Theory** (FST) in the early1960's as a means of modeling uncertainty, vagueness, and imprecision of human natural language. It was built on the basis that as the complexity of a system increases, it becomes more difficult and

eventually impossible to make a precise statement about its behavior, eventually arriving at a point of complexity where the fuzzy logic method born in humans is the only way to get at the problem. **Fuzzy Set Theory** is concerned with application of approximate methods to imprecisely formulated problems, data or real world systems, which are of computational complexity [16]. **Performance** is effectiveness of a system which is assessed or judged. **Transformation** is a process by which one mathematical entity can be derived from one another. **Criteria** are accepted standards used in making decisions or judgments about something.

[12] described *Fuzzy Set Theory (FT)* as the extension of classical set theory. The basic idea is that the membership of a value to a set cannot only assume the two values "yes" or "no", but can be expressed by gradual membership function within a range from zero to normally "1" in case of full membership degree. Membership function can assume several forms, and in practice triangular or trapezium forms are often used (Figure 1).

2. PROBLEM DEFINED

The Collocation extraction techniques used in the paper are of 1) Berry-Roughe, 2) Church and Hanks, 3) Kita, 4) Shimohata, 5) Blaheta and Johnson, and 6) Pearce. These techniques are in rough (imprecise, inexact or fuzzy) ranges, reflecting the variability in how each technique could be implemented and the uncertainties involved in projecting the impacts of the techniques. For a meaningful numerical research, as stated in the introduction, these ordinal ratings need to be transformed to numerical ratings and this forms the thrust of the paper. That is, to transform opinion held by human beings, which would be "fuzzy" (e.g. low, mid-high performance) to being very precise (e.g. 15%, 80% performance), that is not "fuzzy" using fuzzy set theory [12], [13].

3. THEORETICAL FOUNDATION

A fuzzy system is a system whose variable(s) range over states that are approximate. The fuzzy set is usually an interval of real number and the associated variables are linguistic variable such as "most likely", "about", etc. [13]. Appropriate quantization, whose coarseness reflects the limited measurement resolution, is inevitable whenever a variable represents a real-world attribute. Fuzzy logic consists of Fuzzy Operators such as "IF/THEN rules", "AND, OR, and NOT" called the *Zadeh operators* [14].

The Membership Function is a graphical representation of the magnitude of participation of each input. It associates a weighting with each of the inputs that are processed, define functional overlap between inputs, and ultimately determines an output response. Once the functions are inferred, scaled, and combined, they are defuzzified into a crisp output which drives the system. There are different memberships functions associated with each input and output response. Some features of different membership functions are: SHAPE - triangular is common, but bell, trapezoidal, haversine and, exponential have been used also; HEIGHT or magnitude (usually normalized to 1); WIDTH (of the base of function); SHOULDERING; CENTER points (centre of the member and OVERLAP (Figure 1) [15].

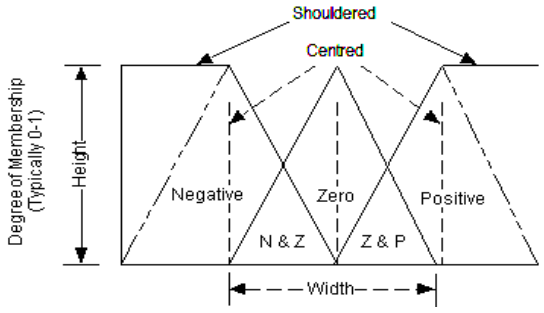


FIGURE 1: Triangular membership function

The Degree of Membership (DOM) is the placement in the transition from 0 to 1 of conditions within a fuzzy set. The degree of membership is determined by plugging the selected input

parameter into the horizontal axis and projecting vertically to the upper boundary of the Membership function(s). Fuzzy Variable includes words like red, blue, good and sweet are fuzzy and can have many shades and tints. A Fuzzy Algorithm is a procedure, usually a computer program, made up of statements relating linguistic variables.

4. METHODOLOGY

The relative effectiveness of these collocation extraction techniques is shown in Table. 1 in terms of four basic criteria: (1) Cost Effectiveness, (2) resolving power, (3) plausibility and (4) Mutual dependency. In the table, assuming, collocation extraction technique of Berry Roughe performs between ***medium to high (m-h)*** on cost effectiveness, ***high (h)*** in terms of resolving power, ***medium to high (m-h)*** on plausibility and ***medium to high (m-h)*** on mutual dependency. Similarly, Church and Hanks technique performs between *medium to high (m-h)* of all the basic criteria. And technique of Kita performs between *low to medium (l-m)* on cost effectiveness and resolving power, *medium (m)* on plausibility, *high (h)* on mutual dependency. Other techniques are also indicated in table 1.

5. NOTATIONS

CRIT	Criteria
CT	Collocation Technique
Coll.	Collocation
CTPER	Collocation Technique Performance
CF	Cost Effectiveness
RP	resolving power
Pl.	Plausibility
MD	Mutual Dependency
m	medium
h	high
l	low

Multi-objective Evaluation of collocation Techniques				
	Ratings on Criteria (high = best)			
Coll. extraction techniques	cf (P)	rp (N)	pl (Q)	md (X)
Berry-Roughe (a)	m-h	h	m-h	m-h
Church and Hanks (b)	m-h	m-h	m-h	m-h
Kita (c)	l-m	l-m	m	h
Shimohata (d)	l-m	l	m	h
Blaheta and Johnson (e)	l-m	l	m	m
Pearce (f)	l-m	l	m	l-m

TABLE 1: Collocation extraction techniques ratings

6. FUZZY VARIABLES

In the paper, the adjectives describing the fuzzy variables and the range of performance are shown in Table 2. The Range of Performance for the individual fuzzy variables is substituted in Table 1 to obtain Table 3.

Fuzzy Variables	Range of Performance %
High(h)	75 – 100
Med-High (m-h)	55 - 80
Medium (m)	35 - 60
Low-Medium(l-m)	15 - 40
Low(l)	0 - 20

TABLE 2: Fuzzy Variables and their ranges.

Multi-objective Evaluation of collocation Techniques				
	Ratings on Criteria (high = best)			
Coll. extraction techniques	cf (P)	rp (N)	pl (Q)	md (X)
Berry-Roughe (a)	55 - 80	75 – 100	55 - 80	55 - 80
Church and Hanks (b)	55 - 80	55 - 80	55 - 80	55 - 80
Kita (c)	15 - 40	15 - 40	35 - 60	75 – 100
Shimohata (d)	15 - 40	0 - 20	35 - 60	75 – 100
Blaheta and Johnson (e)	15 - 40	0 - 20	35 - 60	35 - 60
Pearce (f)	15 - 40	0 - 20	35 - 60	15 - 40

TABLE 3: Fuzzy Range of Performance for the individual fuzzy variables.

7. FUZZY MAPPING

The fuzzy variables in Table 1 were transformed to numerical ratings using *Fuzzy Set Theory* as shown in Figures 2–6.

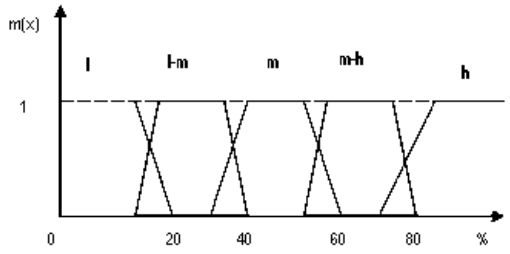


FIGURE 2: Trapezoidal membership function

8. AGGREGATION OF FUZZY SCORES

Using Figure 3, for each Collocation Technique (CT) *i* and each criterion (CRIT) *j*,

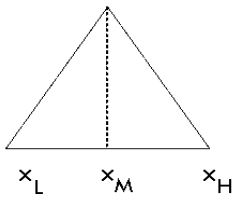


FIGURE 3: Aggregation of Fuzzy Scores.

i = 1, 2, 3, ...6. and *j* = 1, 2, 3, 4.

For CRIT (*j*) when CT (*i*, *j*) = x_L THEN CTPER (*i*, *j*) = L

For CRIT (*j*) when CT (*i*, *j*) = x_M THEN CTPER (*i*, *j*) = M

For CRIT (*j*) when CT (*i*, *j*) = x_H THEN CTPER (*i*, *j*) = H

Where, CRIT (*j*) ≡ Criterion *j* (*j* = 1, 2, 3, 4)

CT (*i*, *j*) ≡ Coll. Techniques *i* under Criterion *j*

CTPER(*i*, *j*) ≡ Coll. Performance Techniques *i* under Criterion *j* Performance

$$\mathrm{CTSCORE(i)} = \sum_{\mathrm{j}} \frac{\mathrm{CT(i,\,j)}}{4} \tag{1}$$

9. MEMBERSHIP FUNCTIONS OF THE FUZZY SETS

Using Aggregation methods for the fuzzy sets to reduce it to a triangular shape for the membership function, overlapping adjacent fuzzy sets were considered with the membership values shown in Figure 4.

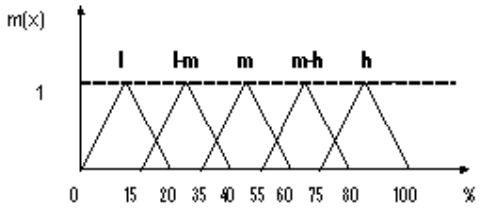
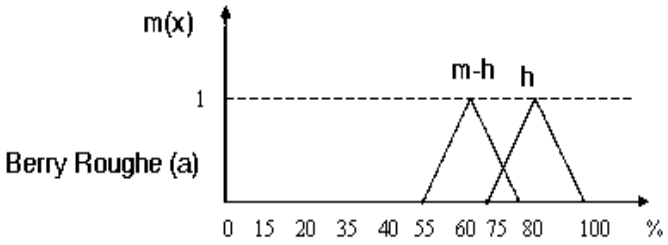
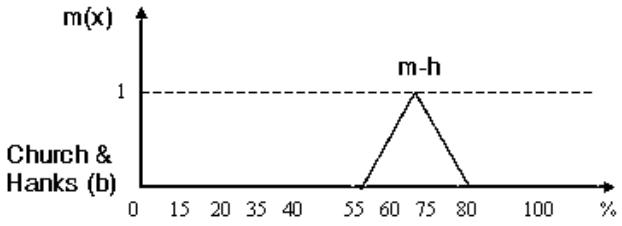


FIGURE 4: Derived Triangular membership function

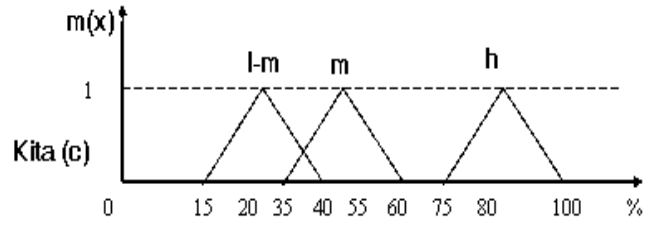
For the techniques and their performances, the membership functions shown in Figure 5 of the fuzzy sets were assigned.



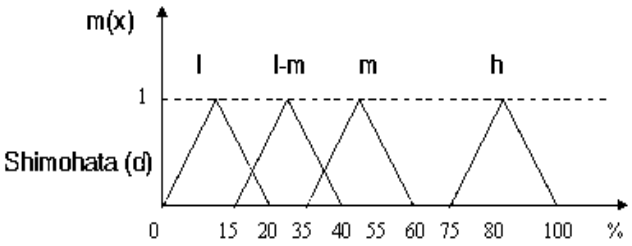
Criteria: (P, Q, X = med-high; N = high)



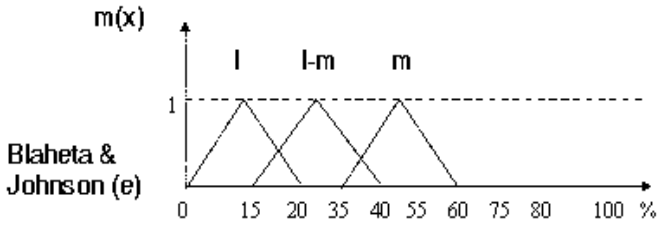
Criteria: (P, N, Q, X = med-high)



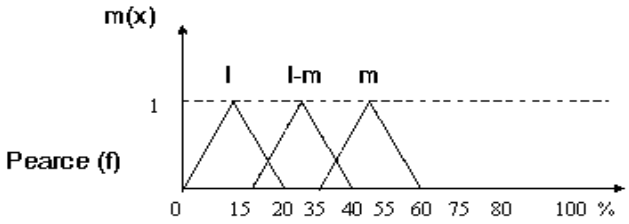
Criteria: (P, N = l-m; Q = m; X = h)



Criteria: (P = l-m; N = l; Q = m; X = h)



Criteria: (P = l-m; N = l; Q, X = m)



Criteria: (P, X = l-m; N= l; Q = m).

FIGURE 5: Derived triangular membership functions for the techniques and the criteria

The ranges in figure 4 and figure 5 were aggregated to singletons. For the average performance of all the techniques, we have the fuzzy scaled rating as shown in Figure 6.

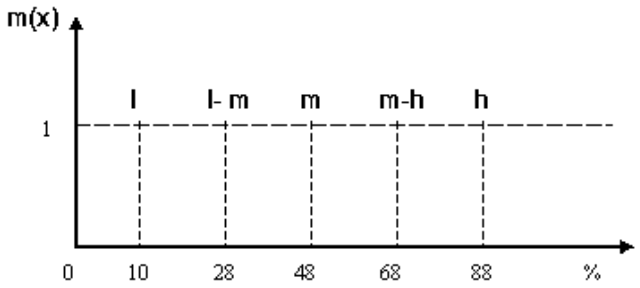


FIGURE 6: Singleton aggregation of the ratings in table 1.

From Figures 2–6, the Membership Values assigned to each set of Universe of Discourse can be tabulated as shown in Table 3.

Coll. extraction techniques	Criteria											
	cf (P)			rp (N)			pl (Q)			md (X)		
Berry-Roughe(a)	m-h			h			m-h			m-h		
	X	Y		X	Y		X	Y		X	Y	
	55	0		75	0		55	0		55	0	
	68	1		88	1		68	1		68	1	
	80	0		100	0		80	0		80	0	
Church and Hanks (b)	m-h			m-h			m-h			m-h		
	X	Y		X	Y		X	Y		X	Y	
	55	0		55	0		55	0		55	0	
	68	1		68	1		68	1		68	1	
	80	0		80	0		80	0		80	0	
Kita (c)	l-m			l-m			m			h		
	X	Y		X	Y		X	Y		X	Y	
	15	0		15	0		35	0		75	0	
	28	1		28	1		48	1		88	1	
	40	0		40	0		60	0		100	0	
Shimohata (d)	l-m			l			m			h		
	X	Y		X	Y		X	Y		X	Y	
	15	0		0	0		35	0		75	0	
	28	1		10	1		48	1		88	1	
	40	0		20	0		60	0		100	0	
Blaheta and Johnson (e)	l-m			l			m			m		
	X	Y		X	Y		X	Y		X	Y	
	15	0		0	0		35	0		35	0	
	28	1		10	1		48	1		48	1	
	40	0		20	0		60	0		60	0	
Pearce (f)	l-m			l			m			l-m		
	X	Y		X	Y		X	Y		X	Y	
	15	0		0	0		35	0		15	0	
	28	1		10	1		48	1		28	1	
	40	0		20	0		60	0		40	0	

TABLE 4: Fuzzy performance ratings of Membership Values assigned to each set of Universe of Discourse.

10. RESULTS

In figure 3 above, for all the collocation extraction techniques, x_L values, x_M values, and x_H values are referred to as *Minimum Performance* (Table 5), *Average Performance* (Table 6) and *Maximum Performance* (Table 7) respectively in the transformation.

10.1 Minimum Performance: The transformed result shows that all the techniques (i.e.of Berry-Roughe, Church and Hanks, Kita, Shimohata, Blaheta and Johnson, and Pearce) have average ratings of 60%, 55%, 35%, 31%, 21% and 16% respectively at Minimum Performance.

Multi-objective Evaluation of collocation Techniques					
	Ratings on Criteria (high = best)				Average Rating on all Criteria
Coll. extraction techniques	cf (P)	rp (N)	pl (Q)	md (X)	
Berry-Roughe (a)	55	75	55	55	60
Church and Hanks (b)	55	55	55	55	55
Kita (c)	15	15	35	75	35
Shimohata (d)	15	0	35	75	31
Blaheta and Johnson (e)	15	0	35	35	21
Pearce (f)	15	0	35	15	16

TABLE 5: Numerical transformation for Minimum Performance

10.2 Average Performance: The transformed result shows that all the techniques (i.e. of Berry-Roughe, Church and Hanks, Kita, Shimohata, Blaheta and Johnson, and Pearce) have average ratings of 73%, 68%, 48%, 43%, 33% and 28% respectively at Average Performance.

Multi-objective Evaluation of collocation Techniques					
	Ratings on Criteria (high = best)				Average Rating on all Criteria
Coll. extraction techniques	cf (P)	rp (N)	pl (Q)	md (X)	
Berry-Roughe (a)	68	88	68	68	73
Church and Hanks (b)	68	68	68	68	68
Kita (c)	28	28	48	88	48
Shimohata (d)	28	10	48	88	43
Blaheta and Johnson (e)	28	10	48	48	33
Pearce (f)	28	10	48	28	28

TABLE 6: Numerical transformation for Average Performance

10.3 Maximum Performance: The transformed result shows that all the techniques (i.e. of Berry-Roughe, Church and Hanks, Kita, Shimohata, Blaheta and Johnson, and Pearce) have average ratings of 85%, 80%, 60%, 55%, 45% and 40% respectively at Maximum Performance.

Multi-objective Evaluation of collocation Techniques					
	Ratings on Criteria (high = best)				Average Rating on all Criteria
Coll. extraction	cf	rp	pl	md	

techniques	(P)	(N)	(Q)	(X)	
Berry-Roughe (a)	80	100	80	80	85
Church and Hanks (b)	80	80	80	80	80
Kita (c)	40	40	60	100	60
Shimohata (d)	40	20	60	100	55
Blaheta and Johnson (e)	40	20	60	60	45
Pearce (f)	40	20	60	40	40

TABLE 7: Numerical transformation for Maximum Performance

11. ANALYSIS:

11.1 Comparison Between the Ordinal Fuzzy Ratings and the Transformed Ratings of all the Different Criteria.

The performance ratings for Collocation Extraction Techniques in terms of cost effectiveness (cf) were fuzzy Table 1, but the performance ratings of all the techniques in terms of cost effectiveness (cf) have been transformed into unique three categories of performances (Minimum, Average, and Maximum) in Table 5-7.

Collocation Extraction Techniques	Ordinal Performance (Fuzzy Ratings)	Minimum Performance (Transformed Ratings)	Average Performance (Transformed Ratings)	Maximum Performance (Transformed Ratings)
Berry Rough	m-h	55	68	80
Church and Hanks	m-h	55	68	80
Kita	l-m	15	28	40
Shimohata	l-m	15	28	40
Blaheta and Johnson	l-m	15	28	40
Pearce	l-m	15	28	40

Similarly, comparisons between the ordinal ratings and the transformed ratings on resolving power (rp), plausibility (pl) and mutual dependency of the criteria can also be shown.

12. CONCLUSION

Using equation (1), we can calculate the Average Scores of different Collocation extraction performance techniques for all the four criteria in respect of x_L referring to as the Minimum Performance, in respect of x_M referring to as the Average Performance, and in respect of x_H referring to as the Maximum performance. Hence their performances ratings can be shown such as $x_L < x_M < x_H$. Fuzzy logic was used to transform ordinal collocation extraction performance ratings that are imprecise and fuzzy in nature to precise and defuzzified numerical ratings that are used in the analysis of performance ratings of different collocation extraction performance techniques. The Technique used is the only way for solving any highly complex problem and can designed its system analysis.

13. REFERENCES

[1] D. Pearce. “A Comparative Evaluation of Collocation Extraction Techniques”. Available: http://www.irec-conf.org/proceedings/irec2002/pdf/169.pdf [Jan. 23, 2011].

[2] A. Thanopoulos, N. Fakotakis, and G. Kokkinakis. “Comparative Evaluation of Collocation Extraction Metrics”. Available: http://www. irec-conf.org/proceedings/irec2002/pdf/128.pdf [Jan.26, 2011]

[3] L.M.Berry-Rogghe. "The computation of collocations and their relevance to lexical studies" in A.J.Aitken, R.W.Balley, and N.Hamilton-Smith, *The Computer and Literacy Studies*, pp.103-112, University Press, Edinburgh, New Delhi, 1973.

[4] K. Ward Church & P. Kanks. "Word association norms, mutual information, and lexicography". *Computational Linguistics*, 16(1):22-29, Mar. 1990.

[5] K. Kita, Y. Kato, T. Omoto, and Y. Yano. "A comparative study of automatic extraction of collocations from corpora: Mutual Information vs. cost criteria". *Journal of Natural Language Processing*, 1(1):21-33, 1994.

[6] S. Shimohata, T. Sugio, and J. Nagata. "Retrieving collocations by co-occurrences and word order constraints". In 35th Conference of the Association for Computational Linguistics (ACL'97),pp 476-481,Madrid, Spain 1997.

[7] F. Smadja. "Retrieving Collocations from Text:Xtract", *Computational Linguistics*, 19(1):143- 177, Mar. 1993.

[8] J.P. Goldman, L. Nerima, and E. Wehril. "Collocation extraction using a syantactic parser", in 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39), pp.61-66, CNRS, Institut de Recherche en Informatique de Toulouse, and Universite des Sciences Sociales, Toulouse, France, Jul., 2001.

[9] D. Lin. "Extracting collocations from text corpora", in First Workshop on Computational Terminology, Montreal, Canada, Aug., 1998.

[10] D. Blaheta and M. Johnson. "Unsupervised learning of multi-word verbs", in 39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39), pp.54-60, CNRS, Institut de Recherche en Informatique de Toulouse, and Universite des Sciences Sociales, Toulouse, France, Jul., 2001.

[11] D. Pearce. "Synonymy in collocation extraction", in NACCL 2001 Workshop: WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Carnegie Mellon University, Pittsburgh, Jun., 2001.

[12] L.A. Zadeh. "Fuzzy sets". *Information and Control*, 8, pp. 338 – 353, 1965.

[13] L.A. Zadeh. "Toward a theory of fuzzy information granulation and its Centrality in human reasoning and Fuzzy logic". *International Journal of Soft Computing and Intelligence*, 90, 2, pp. 111 – 127, 1997.

[14] T. Sowell. *Fuzzy-Logic*. [Online]. Available : http://www.fuzzy logic.com/ch3.htm, 2005 [Jan.15,2011].

[15] S.D.. Kaehler: *Fuzzy Logic*.[Online]. Available : http://www.seattlerobotics.org/ encoder/mar98/fuz /flindex.html, (1998) [Jan.28, 2011].

[16] After Reviewing:

[17] E.A., Shyllon."Techniques for Modelling Uncertainties Inherent in Geomatics Data", *First International Symposium on Robust Statistics and Fuzzy Technics in Geodesy and GIS*, Zurich: Swiss Federal Institute of Technology Zurich (ETH), Institute of Geodesy and Photogrammetry, pp. 139-143, 2001.