Fuzzy Set Theoretic Approach to Collocation Extraction

Raj Kishor Bisht
Dept. of Computer Science
Amrapali Institute of Management
and Computer Application,
Haldwani (Uttarakhand)-India

H. S. Dhami
Dept. of Mathematics
University of Kumaun,
S.S.J.Campus Almora
(Uttarakhand) -India

ABSTRACT
Fuzzy approach deals with the linguistic properties of elements such as beauty, coldness, hotness etc. Collocations are linguistically motivated. Decision of word combination for being collocation is a linguistic term as merely co-occurrence of word combinations does not signify the presence of collocation. Thus collocation extraction can be made possible by looking its linguistic aspect. In the present paper, an attempt has been made to make two different fuzzy sets of word combinations to be considered for collocations. Mutual information and t-test have been taken as basis for the construction of fuzzy sets. Two fuzzy set theoretical models have been proposed to identify collocations. It has been shown that fuzzy set theoretical approach works very well for collocation extraction. The working data has been based on a corpus of about one million words contained in different novels constituting project Gutenberg available on www.gutenberg.org.

General Terms
Natural Language processing, Computational lexicography

Keywords
Collocation, Fuzzy set, Mutual Information, t-test

1. INTRODUCTION
'Collocations' are a class of word groups which lie between idioms and free word combinations. However, it is typical to draw a line between a phrase and a collocation. Idioms and phrases may be defined as an expression in the language that is peculiar to itself. It becomes well nigh impossible to guess the meaning of an idiom from the word it contains (e.g. At the eleventh hour). And, moreover, the meanings that idioms have are often stronger than the meanings of non-idiomatic phrases. For instance, 'look daggers at someone' is more emphatic than 'look angrily at someone', although both of them have the same meaning [5]. On the other hand, in a free word combination, a word can be replaced by another word without seriously modifying the overall meaning of the composite unit so that one can not easily predict it from the remaining ones. For example, 'end of the day' can not be predicted from 'end of the lecture', if we replace 'day' by 'lecture'. According to Kathleen R. McKeown and Dragomir R. Radev [8] 'collocations are arbitrary, language specific, recurrent in context and common in technical language'. Collocations are utilized for many natural language applications such as, machine translation, computational lexicography, information retrieval, natural language generation etc [10]. Collocation translation improves the quality of machine translation. For example, 'public opinion' in English is '*janata ki raay'* in Hindi, 'pocket money' in English is '*jeb kharch'* in Hindi. Automatic identification of important collocations to be listed in a dictionary is the task of computational lexicography. Adequate knowledge of collocations can improve the performance of information retrieval system.

Statistical methods have shown a remarkable presence in collocation extraction. Frequency measure was used by Choueka et al [2] to identify a particular type of collocations. Church and Hanks [3] used mutual information to extract word pairs that tend to co-occur within a fixed size window (normally 5 words), in which extracted words may not be directly related. Smadja [11] extracted collocations through a multi-stage-process taking the relative positions of co-occurring words into account. Church and Gale [4] used the χ^2 - test for the identification of translation pairs in aligned corpora. Collocations extraction and their use in finding word similarity was suggested by Dekang Lin [9]. The use of t-test to find words whose co-occurrence patterns best distinguish between two words was suggested by Church and Hanks [3]. Dunning [6] applied likelihood ratio test to collocation discovery. Marc Weeber et al [16] devised an extraction system for the full word frequency ranges which computes the significance of association by the log-likelihood ratio and Fisher's exact test. Diana Zaiu Inkpen and Graeme Hirst [15] extended a lexical knowledge-base of near-synonym differences with knowledge about their collocational behaviour. Pavel Pecina [13] made an empirical evaluation of a comprehensive list of automatic collocation extraction methods using precision-recall measures. Violeta Seretan and Eric Wehrli [14] pointed out several language-specific issues related to extraction and proposed a strategy for coping with them. Afsaneh Fazly and Suzanne Stevenson [12] identified several classes of multiword expressions.

Almost all the techniques of collocation extraction look at whether the probability of seeing a combination differs significantly from what we would expect from their component words and reject those word combinations that do not. To decide whether a word combination makes a collocation or not is a vague measurement. One can not apply a particular rule of collocation extraction for every word combinations; thus fuzzy approach is quite useful for collocation extraction. In a classical or crisp set we assign only two values 0 or 1 to different elements depend on their belongingness to the set. If an element is a member of the set then it is 1 otherwise 0. This approach is well

defined for exact properties, such as for the set of positive real numbers on a set of real numbers, we may assign 1 for every positive real number and 0 for every non positive real number but this approach does not work well for linguistic terms such as good student, hot temperature etc. because no well defined definition is there for such terms. Instead of assigning 0 or 1, we use the closed interval [0, 1]. For a fuzzy set, we define a grade of membership for each element which shows its degree of belongingness to the set. Zero grade of membership indicates that the element does not belong to the set and one grade of membership gives full support to the element for its belongingness to the set. We can assume a set of collocations in which every word combination is a member of the set with different grades of membership. In the present paper we have made an attempt to find out the membership function for word combinations by utilizing two previous approaches of collocation extraction, that is, mutual information and t-score. The working data has been based on 1 million words corpus compiled by taking some of the novels contained in project Gutenberg available at www.gutenberg.org/etext/<no.> (See appendix).

The structure of the paper is as follows: In Section 2, we have defined two fuzzy sets obtained by fuzzyfying the mutual information scores and t-scores. In section 3, fuzzy set theoretical model for collocation extraction has been proposed. Evaluation of the proposed model will be the part of section 4. Finally, Section 5 deals with the discussion and conclusions on the present study.

2. FUZZYFICATION OF COLLOCATION EXTRACTION TECHNIQUES

In this section, we have mentioned the two techniques of collocation extraction, that is, mutual information and t-score. We have found the grade of membership for each word combination based on these two methods.

2.1 Mutual Information

Mutual information from information theory has been utilized to find the closeness between word pairs by Church & Hanks [3]. Mutual information for two events x and y is defined as:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x).P(y)} \quad (1)$$

If we write w_1 and w_2 for the first and second word respectively, instead of x and y, then the mutual information for the two words w_1 and w_2 is given by:

$$I(w_1,w_2) = \log_2 \frac{P(w_1,w_2)}{P(w_1).P(w_2)} \quad (2)$$

where $P(w_1,w_2)$ is the probability of two words w_1 and w_2 coming together in a certain text and $P(w_1)$ and $P(w_2)$ are the probabilities of w_1 and w_2 appearing separately in the text, respectively.

If $P(w_1,w_2) = P(w_1).P(w_2)$, that is, the two words are independent to each other, then $I(w_1,w_2) = 0$, which indicates that these two words are not good candidates for collocation. A high mutual information score signifies the presence of a collocation.

2.1.1 Fuzzification of Mutual Information

The term 'high mutual information score' is quite vague as it does not provide a basis to say which mutual information score can be considered as high. Let us consider a fuzzy set A of collocations, then each word combination will be a member of the set A with a particular grade of membership. Grade of membership can be defined with the help of mutual information scores. To find the grade of membership using the mutual information scores, we have analyzed the pattern of mutual information scores. Instead of looking for every mutual information score, we have classified mutual information scores into small class intervals and assigned ranks to word combinations according to the classes. So that word combinations falling under a class have same grade of membership. Table 1 shows the mutual information and their corresponding ranks.

Table 1: Mutual information scores and corresponding ranks.

Mutual Information score	Rank
0 - 0.25	1
0.26 - 0.50	2
0.51 - 0.75	3
0.76 - 1.00	4
……	…
……	…
12.76 - 13.00	51
……	…

Class intervals have been taken too small as we have found that small class interval leads to accuracy in the prediction of ranks from the distribution. We can define a function to get the rank of a word combination from its mutual information. Let $x \in R^+$ be the mutual information of a word combination and $y \in I^+$ be the corresponding rank, then $f : R^+ \to I^+$ such that $y = f(x) = \left\lceil \frac{x}{0.25} \right\rceil$. $\lceil x \rceil$ is the ceiling function.

To find the grade of membership from the rank, we can define an another function $g: I^{+} \rightarrow [0,1]$ such that $y = g(x)$, where $x \in I^{+}$ is the rank and $y \in [0,1]$ is the corresponding grade of membership. For defining the function $g(x)$, we know that it should tend to zero when x tends to one (lowest mutual information rank) and one when x tends to infinity (highest mutual information score) respectively. Therefore we can define $g(x)$ as follows:

$$y = g(x) = \frac{\log x}{\log(x+10)} \tag{3}$$

From the function $y = g(x)$, it is clear that when $x \rightarrow 1,\ y \rightarrow 0$ and $x \rightarrow \infty,\ y \rightarrow 1$.

Finally grade of membership for a word combination based on the mutual information scores can be given as:

$$A_I = g\ of\ (x) = g(f(x)) \tag{4}$$

We have taken the example of eighty word pairs from the compiled corpus. Table 2 shows the mutual information scores and their corresponding grades of membership for different word combinations.

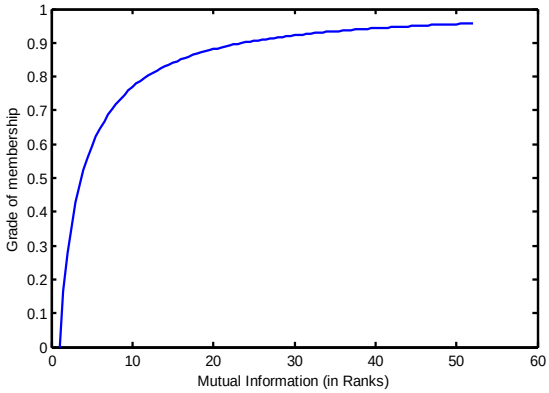

Fig.1: Grade of membership for mutual information scores

2.2 The t-test

The t-test has been used by Church & Hanks [3] for collocation discovery to test the validity of a hypothesis. For that purpose, we formulate a null hypothesis H_0 that the two words w_1 and w_2 appear independently in the text. So under the null hypothesis H_0, the probability that the words w_1 and w_2 are coming together is simply given by:

$$P(w_1, w_2) = P(w_1).P(w_2).$$

The null hypothesis has been tested by using t-test. If the null hypothesis is accepted, we conclude that the occurrence of two words is independent of each other. Otherwise, we may conclude that they depend on each other, that is, they form collocations. In t-test we use the null hypothesis that the sample is drawn from a distribution with mean μ, taking sample mean and variance into account. The t-test considers the difference between the observed and expected mean. The t statistic is defined as:

$$t = \frac{\bar{x} - \mu}{\sqrt{s^2 / N}} \sim t_{n-1}(\alpha) \tag{5}$$

where $\bar{x}$ is the sample mean, s^2 is the sample variance, N is the sample size, μ is the mean of the distribution and $t_{n-1}(\alpha)$ denotes a t- distribution with (n-1) degrees of freedom at α level of significance. To apply t- test for testing the independence of two words w_1 and w_2, we assume that $f(w_1)$, $f(w_2)$ and $f(w_1, w_2)$ are the respective frequencies of the word w_1, w_2 and $w_1\ w_2$ in the corpus and N is the total number of words / bigrams in the corpus. Then, we have,

$$P(w_1) = \frac{f(w_1)}{N} \text{ (say } p_1), \quad P(w_2) = \frac{f(w_2)}{N} \text{ (say } p_2),$$

$$P(w_1, w_2) = \frac{f(w_1, w_2)}{N} \text{ (say } p_{12}),$$

The null hypothesis is

$$H_0 : P(w_1, w_2) = P(w_1).P(w_2) = p_1 . p_2$$

If we select bigrams (word pairs) randomly then the process of randomly generating bigrams of words and assigning 1 to the outcome that the particular word combination for which we are looking for is a collocation and 0 to any other outcome follows a Bernoulli distribution. For the Bernoulli distribution we have

Mean $(\mu) = p$ and Variance $(\sigma^2) = p(1-p)$.

Thus, if the null hypothesis is true, the mean of the distribution is $\mu = p_1.p_2$. Also, for the sample, we have $P(w_1, w_2) = p_{12}$. Therefore, using Binomial distribution, sample mean $\bar{x} = p_{12}$ and sample variance $s^2 = p_{12}(1-p_{12})$. Using (5), we calculate the value of $|t|$ and compare it with the tabulated value at given level of significance. If the value of $|t|$ for a particular bigram is greater than the value obtained from the table, we reject the null hypothesis, which indicates that the bigram may be

considered as a collocation. We have chosen the level of significance $\alpha = .005$ for which t = 2.57.

2.2.1. Fuzzification of t-score

To accept only those bigrams for collocations which have $|t|$ score greater than 2.57 is accurate as far as we take the t-test into consideration but extraction of collocation is not a pure mathematical job since the decision as to what constitutes a collocation is affected by its linguistic aspect also. This provides us a reason to think about those word combinations whose t scores are less than 2.57 but very close to it. Therefore it opens a way to make a fuzzy set for collocations based on t scores. The membership function for a word combination x can be defined as follows:

$$A_t(x) = \begin{cases} 1 & \text{if} \quad |t| \geq 2.57 \\ \dfrac{|t|}{2.57} & \text{otherwise} \end{cases} \qquad (6)$$

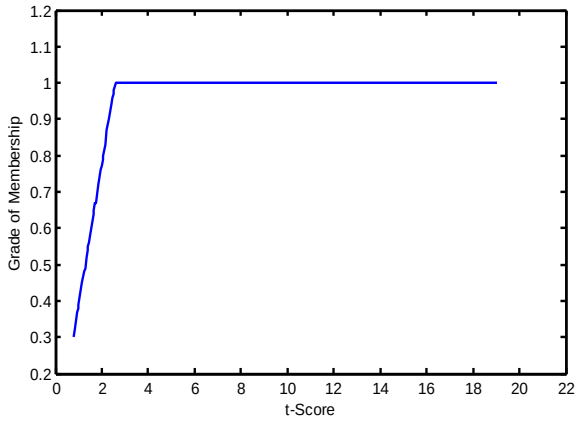

Fig.2: Grade of membership for collocation using t-score

3. FUZZY DECISION MAKING FOR COLLOCATION EXTRACTION

In this section, we have presented fuzzy set theoretical models for collocation extraction. In this model different opinions have been aggregated to find the fuzzy set of collocation. We have used the most common method based on the probabilistic interpretation of membership functions given by multiple experts [7]. Different experts are asked to for some $x \in X$ to valuate its belongingness to A, where A is a fuzzy set on X that represent a linguistic term associated with a given linguistic variable. For some $x \in X$, let $A_r(x)$ denote the answer of the r^{th} expert $(r \in N)$ in the term of belongingness of x to A, where $A_r(x)$ has only two values 0 and 1 for $x \in A$ and $x \notin A$ respectively. Then the membership function can be defined as follows:

$$A(x) = \frac{\sum_{r=1}^{n} A_r(x)}{n}$$, where n is the number of experts.

Here, for a word combination x, $A_I(x)$ and $A_t(x)$ be the opinions of the mutual information score and t- score respectively, in terms of grade of membership of x to its belongingness to A. For a word combination x, The grade of membership can be given as $A(x) = \dfrac{A_I(x) + A_t(x)}{2}$.

On using the above fuzzy decision model, we can obtain the grade of membership of each word combination for being collocation. A word combination that attains a maximum grade of membership can be taken as collocation.

4. EVALUATION OF FUZZY SET THEORETIC APPROACH

For evaluating the fuzzy set theoretic approach, we have taken the example of eighty word pairs from the compiled corpus. We have calculated the mutual information scores and t-scores for different word combinations and also their corresponding grades of membership. Table 2 shows the frequencies $f(w_1)$, $f(w_2)$, $f(w_1 w_2)$ of words w_1, w_2 and their combination $w_1 w_2$ respectively with their respective mutual information score, t-score and grades of membership. Table 2 shows the grades of membership for different bigrams in the text using the fuzzy set theoretic model. We can choose different standards (grades of membership) for collocations extraction as values near to 1 show a high grade of membership for collocation. Validity of the word pairs given in table 2 have been checked through www.thefreedictionary.com, Cambridge Advanced learner's Dictionary and English to Hindi Translation point of view. Only 17 word pairs (star marked in table 2) have been found meaningful as collocations.

If we look at the translation of the some of the word pairs from English to Hindi we found that the words marked as asterisk in table 2 form meaningful combination. "Christmas eve" is translated as "Christmas ki purv sandhya", "Public opinion' is translated as "Janta ki raay" or "janmat", "human being" as "manav" or "manushya" , "young man" as "Naujawan" , "human nature" as "Manav Swabhaw" , "take care" as "kyayal rakhana" .

On the basis of this, we can compare the results of mutual information and t-score with the results obtained by the proposed model. The mutual information does not provide a criterion for collocation extraction except saying high mutual information score shows the presence of a collocation. Precision and recall will depend upon the choice of the high mutual information score. However we can take different criteria of mutual information for calculating precision and recall. Table 3 shows the precision and recall for different mutual information scores. For t-score precision is 46% and recall is 70%. Table 4 shows the precision and recall of the proposed models.

Table 2: Mutual Information, t-score and the respective grades of membership for different word combinations.

W_1	W_2	$f(w_1)$	$f(w_2)$	$f(w_1w_2)$	MI	A_i	$\|t\|$	A_t	A
*Christmas	eve	72	33	9	11.96	0.95	3.00	1.00	0.98
*base	camp	54	55	7	11.27	0.95	2.64	1.00	0.98
*public	opinion	190	103	10	9.07	0.94	3.16	1.00	0.97
*both	sides	409	69	11	8.68	0.93	3.31	1.00	0.97
*human	being	182	735	30	7.88	0.93	5.45	1.00	0.97
*great	deal	911	118	20	7.61	0.92	4.45	1.00	0.96
*human	nature	182	251	7	7.33	0.92	2.63	1.00	0.96
strong	enough	172	657	13	6.92	0.91	3.58	1.00	0.96
*more	than	2124	1563	369	6.87	0.91	19.05	1.00	0.96
*take	care	808	228	20	6.83	0.91	4.43	1.00	0.96
*young	man	741	2138	147	6.61	0.91	12.00	1.00	0.96
*early	days	182	497	7	6.35	0.91	2.61	1.00	0.96
long	journey	967	100	7	6.25	0.91	2.61	1.00	0.96
last	night	846	856	47	6.09	0.90	6.76	1.00	0.95
*fire	bucket	291	15	5	10.23	0.94	2.23**	0.87	0.91
away	from	862	3945	135	5.38	0.89	11.34	1.00	0.95
came	along	1360	367	13	4.77	0.88	3.47	1.00	0.94
every	night	676	856	13	4.56	0.87	3.45	1.00	0.94
trench	life	99	1102	6	5.85	0.90	2.41**	0.94	0.92
night	before	856	1164	18	4.25	0.86	4.02	1.00	0.93
must	take	1144	808	16	4.18	0.86	3.78	1.00	0.93
last	time	846	1463	20	4.09	0.85	4.21	1.00	0.93
strong	man	172	2183	7	4.29	0.86	2.51**	0.98	0.92
long	after	967	1304	18	3.91	0.85	3.96	1.00	0.93
might	even	1143	768	10	3.58	0.83	2.90	1.00	0.92
come	over	1358	1394	19	3.40	0.83	3.95	1.00	0.92
*look	upon	756	1913	12	3.12	0.81	3.07	1.00	0.91
little	episode	1630	16	4	7.33	0.92	1.99**	0.77	0.85
almost	every	518	676	6	4.17	0.86	2.31**	0.90	0.88
*evil	eye	124	259	4	7.03	0.92	1.98**	0.77	0.85
time	before	1463	1164	11	2.76	0.79	2.83	1.00	0.90
long	way	967	1084	8	3.00	0.80	2.48*	0.96	0.88
painful	experience	27	87	3	10.39	0.95	1.73**	0.67	0.81
make	use	963	222	5	4.62	0.87	2.15**	0.83	0.85
very	like	1410	1602	11	2.36	0.76	2.67	1.00	0.88
dark	shadow	279	93	3	6.92	0.92	1.72**	0.67	0.80
last	century	846	32	3	6.86	0.91	1.72**	0.67	0.79
like	some	1602	1786	11	2.01	0.72	2.50**	0.97	0.85
cheerful	noise	48	720	3	6.51	0.91	1.71**	0.67	0.79
night	air	856	475	5	3.69	0.84	2.06**	0.80	0.82

W_1	W_2	$f(w_1)$	$f(w_2)$	$f(w_1 w_2)$	MI	A_i	$\lvert t \rvert$	A_t	A
only	because	1187	371	5	3.58	0.83	2.05**	0.80	0.82
your	book	2888	153	5	3.57	0.83	2.05**	0.80	0.82
another	half	693	696	5	3.45	0.83	2.03**	0.79	0.81
welcome	relief	70	71	2	8.72	0.93	1.41**	0.55	0.74
*national	guard	39	163	2	8.37	0.93	1.41**	0.55	0.74
good	terms	1299	88	3	4.79	0.88	1.67**	0.65	0.77
horrible	thing	55	580	2	6.04	0.90	1.39**	0.54	0.72
stark	madness	6	22	1	12.96	0.96	1.00**	0.39	0.68
usual	hour	115	344	2	5.73	0.90	1.39**	0.54	0.72
step	towards	135	348	2	5.48	0.89	1.38**	0.54	0.72
away	down	862	1517	6	2.27	0.75	1.94**	0.76	0.76
valid	reason	6	151	1	10.18	0.94	1.00**	0.39	0.67
might	still	1143	799	5	2.52	0.77	1.85**	0.72	0.75
spiritual	creature	15	62	1	10.14	0.94	1.00**	0.39	0.67
rapid	motion	32	42	1	9.61	0.94	1.00**	0.39	0.67
clumsy	fashion	11	126	1	9.57	0.94	1.00**	0.39	0.67
like	myself	1602	372	4	2.82	0.79	1.72**	0.67	0.73
visible	effort	27	73	1	9.06	0.94	1.00**	0.39	0.67
empty	tent	128	16	1	9.00	0.94	1.00**	0.39	0.67
huge	space	35	64	1	8.87	0.93	1.00**	0.39	0.66
peasant	girl	10	253	1	8.70	0.93	1.00**	0.39	0.66
wild	dreams	136	46	1	7.39	0.92	0.99**	0.39	0.66
except	myself	81	1602	2	4.02	0.85	1.33**	0.52	0.69
wrong	way	121	1084	2	4.00	0.85	1.33**	0.52	0.69
round	upon	387	1913	4	2.51	0.77	1.65**	0.64	0.71
last	link	864	15	1	6.34	0.91	0.99**	0.38	0.65
looking	through	431	1010	3	2.86	0.79	1.49**	0.58	0.69
human	affairs	182	86	1	6.07	0.90	0.99**	0.38	0.64
*water	level	286	65	1	5.82	0.90	0.98**	0.38	0.64
most	powerful	723	33	1	5.46	0.89	0.98**	0.38	0.64
along	over	367	1394	3	2.62	0.78	1.45**	0.56	0.67
little	chap	1630	132	2	3.29	0.82	1.27**	0.49	0.66
weary	night	46	856	1	4.74	0.87	0.96**	0.37	0.62
great	emotions	911	45	1	4.68	0.87	0.96**	0.37	0.62
things	behind	622	415	2	3.03	0.81	1.24**	0.48	0.65
very	dark	1410	279	2	2.42	0.76	1.15**	0.45	0.61
right	about	757	1723	4	1.69	0.68	1.38**	0.54	0.61
*make	sense	963	192	1	2.51	0.77	0.82**	0.32	0.55
only	chance	1187	200	1	2.15	0.74	0.77**	0.30	0.52
little	after	1630	1304	4	0.98	0.52	0.99**	0.38	0.45

*actual collocation ** Rejected for collocation $(|t| < 2.57)$

Table 3: Precision and recall for Mutual Information.

Mutual Information (equal to or more than)	Precision (in %)	Recall (in %)
10.0	43	18
8.0	35	35
6.0	39	82
4.0	27	88

Table 4: Precision and recall for fuzzy decision model

Grade of membership (equal to or more than)	Precision (in %)	Recall (in %)
0.98	100	12
0.95	**79**	**65**
0.90	48	76
0.85	38	76

From table 3 and 4, we can observe that fuzzy set theoretical model based on mutual information and t-score provides a better opportunity to extract collocations than using mutual information and t-score alone. Particularly, at more than or equal to .95 grade of membership the model has shown a good result.

5. DISCUSSION

The present work was carried out to utilize the fuzzy approach for collocation extraction. We have calculated the mutual information scores and t-scores for different word combinations. We have found that these two methods alone are not sufficient to extract collocations; however these methods form a strong basis for collocation extraction. In mutual information score it is tough to decide which score can be considered as high score and in t-score the values less than but near to the chosen value of 't' may have the capability of making collocations. These points have opened the way to think in the direction of utilizing the fuzzy set theoretic approach and we have fuzzified both the techniques. Results prove the utility of fuzzy approach for collocation extraction. Therefore we conclude that the proposed model based on fuzzy set theoretical approach opens a new dimension for collocation extraction.

5. REFERENCES

[1] Bellman, R.E., Zadeh, L. A. (1970). Decision making in fuzzy environment. *Management Science* 17(4) 141-164.

[2] Choueka, Y., Klien, T., Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal for Literary and Linguistic computing* Vol 4, 34-38.

[3] Church Kenneth W., Hanks, Patrick. (1989). Word association norms, mutual information and lexicography. *In Proceedings of the 27th meeting of the Association of Computational Linguistics* 76-83.

[4] Church, Kenneth W., Gale, William A. (1991). Concordance for parallel text. *In proceedings of the seventh annual conference of the UW centre for new OED and text research* Oxford 40-62.

[5] Cambridge International Dictionary of Idioms (1998). UK, CUP.

[6] Dunning, Ted. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. Vol 19 61-74.

[7] Klir, George J. & Yuan Bo. (2001). *Fuzzy sets and fuzzy logic theory and application* Prentice Hall of India

[8] Kathleen R. McKeown and Dragomir R. Radev . (2000). *Online manual.* Available at:

http://citeseer.ist.psu.edu/mckeown00collocations.html

[9] Lin, Dekang. (1998). Extracting collocations from text corpora. *In first workshop on Computational terminology*, Montreal, Canada.

[10] Manning, Christopher D., Schutze Heinrich. (2002). *Foundations of Statistical Natural Language Processing*, MIT Press.

[11] Smadja, Frank.(1993). Retrieving collocations from text: Xtract. *Computational Linguistic* Vol 19(1) 143-177.

[12] Fazly, A. , Suzanne S. (2007). "Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures". *In Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, 9–16,

[13] Pecina Pavel. (2005). "An Extensive Empirical Study of Collocation Extraction Methods". In *Proceedings of the ACL Student Research Workshop*, 13–18,

[14] Seretan V. , Wehrli E. (2006). "Multilingual Collocation Extraction: Issues and Solutions". *In Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, 40–49,

[15] Inkpen Diana Zaiu, Hirst Graeme. (2002). "Acquiring Collocations for Lexical Choice between Near-Synonyms" *In Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon*. 67-76.

[16] Weeber Marc, Vos Rein. (2000). "Extracting the Lowest-Frequency Words: Pitfalls and Possibilities". *Computational Linguistics* Volume 26, Number 3, 301-317.