Journal of Computer Science 1 (3): 400-404, 2005
ISSN 1549-3636

Semantic Refinement of Multi-meaning Words by T-norms

F. Sadeghi, H. Khosravi and E. Eslami
Faculty of Mathematics and Computer Science
Shahid Bahonar University of Kerman, Kerman, Iran
International Center for Science and High Technology and Environment Sciences, Kerman, Iran

Abstract: In this paper we consider a multi-meaning word which appears in a sentence together with some connected words. To refine the meaning of this multi-meaning word, we propose a method based on T-norms which are reasonable mathematical models of logical conjunction operator in fuzzy logic. The preliminary results derived from a prepared corpus are encouraging.

Key words: Multi-Meaning Word, NLU, NLP, Ambiguity, Connected Word, Correct Meaning, T-Norm,
AMS Subject Classification: 03B52, 03B65, 08T50

INTRODUCTION

One of the most important problems in Natural Language Understanding (NLU) is the misunderstanding of the real meaning of a multi-meaning word. Multi-meaning words are considered as part of ambiguity in languages. Natural Language (NL) is presented into two different general types: spoken and written language [3]. Ambiguity in the first one happens when at least two words have the *same* pronunciation but different written forms and of course different meanings. For example in English ([tü]: {to, too, two}) and in Farsi words ([ord], " ": {ate) (In this paper we follow the pattern as the first word is the pronunciation of the word in Farsi. The second is written in Farsi, and the last is its meanings in English) and ([ord] , " "{small}). Another type of ambiguity in this category happens when words with the same pronunciation and the same written forms have different meanings. For example in English ("saw": sÔ":{past of see, a cutting tool, to cut, maxim}) and in Farsi the word ([šir], " " : {milk, lion, tap}). In fact the word ([šir], " ") in Farsi has three different meanings as: 1) White fluid secreted by female mammals as food for their young children , 2) A wild animal of the cat family and 3) Tubular plug with a device for allowing liquid to flow through which are named in English by milk , lion and tap respectively. We can use the similar procedure to choose different names or notations from another language to distinguish the meaning of a multi-meaning word to save time and space.

Ambiguity in written languages happens when words with the same written form have different meanings. There are two types: words with same pronunciation like ([šir], " ": {milk, lion, tap}) and words with different pronunciation like ([lem], " ": {scientist}) and ([lam] , " ": {world}). In this paper we focus on ambiguity in written words with same pronunciation, same written forms, but different meanings.

There are different methods approaching this matter. For example [1, 5] which are based on logic and [2, 4, 8] using WordNet and meaning evaluation on probable logic. But we are using T-norms as appropriate models for conjunction connective in fuzzy logic to refine the multi-meaning word. We review different T-norms in section 4, applying them as will be described in section 5 and choose the best one to get natural results as expected.

MATERIAL AND METHODS

This section describes basic definitions and the notation used in this paper. We consider w as a multi- meaning word with w_1, w_2,…,w_n presenting w's different meanings. In addition a_1, a_2,…,a_k are *connected* words to w. The word a is connected to w, when seeing w and a together in a sentence brings a precise meaning of w to memory. The amount of clarity made by each connected word for any meaning of a multi-meaning word is called the degree of connection which is a number between zero and one. Although all the words in a natural language could be connected to each other with some degree of connection, we consider the connected words as those words which have a reasonable degree of connection. The sentence mentioned in this paper, is supposed to be easy and clearly understandable sentence.

For this research we selected five multi-meaning words from Farsi. They are ([šir] " : {milk, lion, tap}), ([zab n], " {language, tongue}), ([b z] ," " {open, falcon, alkaline}), ([t r], " ", {fourth month of the Persian year, bullet}) and ([t r]," ", {blear, web, ukulele}). For the purpose of this research a corpus of 7800 sentences was collected from different people in different situations. We extracted the connected words

to the five multi-meaning words mentioned above by consulting expert people.

Let w be one of these five multi-meaning words and $A=\{a_1, a_2,\ldots,a_k\}$ be the set of connected words to w. Consulting expert people leads us to partition A into classes $[c_1]$, $[c_2]$,…$[c_t]$ such that each $[c_i]$ is a subset of A whose members have the same degree of connection to all meanings of w. To calculate the degree of connection of each class to all different meanings of a multi-meaning word w, we consider the following two criteria: 1) The number of times it has occurred in the corpus, and 2) Consulting expert people. This information could be summarized in a t×n table CDCCW-Table called "Connection Degree of Classified Connected Words (CDCCW)". In this table $R_{ij} = R([c_i], w_j)$ means the degree of connection of the words in the i^{th} class to the j^{th} meaning.

CDCCW-Table

	w_1	w_2	…	w_n
$[c_1]$,	R_{11}	R_{12}	…	R_{1n}
$[c_2]$	R_{21}	R_{22}	…	R_{2n}
⋮	⋮	⋮	⋮	⋮
$[c_t]$	R_{t1}	R_{t2}	…	R_{tn}

MATHEMATICAL MODELING OF THE PROBLEM

Since we are dealing with a vague situation and like to make decisions in a no exact environment, in this section we make a mathematical modeling of the main problem. In fact we are facing the following problem: Let w be a multi-meaning word with the following meanings such as say w_1, w_2,…w_n and suppose w appears in a sentence in which some words a_{i1},… a_{im} from the set of all connected words $A = \{a_1,\ldots a_k\}$ also exist. Our main problem is finding the correct meaning of w in this sentence. Or when we translate the sentence containing the multi-meaning word w, which w_j is the best meaning for translating w to another language? To answer the above question, consider the following sentences:

$\varphi_{1j} :=$ a_{i_1} is connected to w_j,

$\varphi_{2j} :=$ a_{i_2} is connected to w_j,

⋮

$\varphi_{mj} :=$ a_{i_m} is connected to w_j, for j=1,…,n , and

$\varphi_j := (a_{i_1} \,\&\, a_{i_2} \,\&\, \cdots \,\&\, a_{i_m})$ is connected to w_j We set, $\varphi_j \equiv \varphi_{1j} \,\&\, \varphi_{2j} \,\&\, \ldots \,\&\, \varphi_{mj}$ In which "&" means logical connective "*and*" , and "≡" means " equivalence " or "if and only if " , i.e. , the truth value of φ_j , $\vartheta(\varphi_j) = \vartheta(\varphi_{1j},\ldots,\varphi_{mj})$. To evaluate the right hand side we model "&" by a T-norm such that

$$\vartheta(\varphi_j) = T(\vartheta(\varphi_{1j}), \vartheta(\varphi_{2j}), \ldots, \vartheta(\varphi_{mj})) .$$

To find the correct meaning w we find each $\vartheta(\varphi_j)$, $j = 1,\ldots,n$ and compare them. Now the question is, which T-norm is better? To answer this question first we review some well-known T-norms and their properties in the next section

T-NORMS

In the following sentences we give the definitions of the function T-norm and introduce some T-norms and their properties. Further, the concept of the negation function is introduced. Concerning the selection of special instances of T-norms, we refer to our core study in the next section.

Definition 4.1: A T-norm (triangular norm) is a function $T:[0,1]\times[0,1]\to[0,1]$ such that for all $a,b,c,d \in [0,1]$

$(T1)$ $T(a,1) = a$ (boundary condition)

$(T2)$ $T(a,b) \le T(c,d)$,if $a \le c$ and $b \le d$ (monotonic property)

$(T3)$ $T(a,b) = T(b,a)$

$(T3)$ $T(a,b) = T(b,a)$ (commutative property)

$(T4)$ $T(a,T(b,c)) = T(T(a,b),c)$ (associative property)

Among a very long list of T-norms, the following six functions are the most important:

$$T_w(x,y) = \begin{cases} x & if \quad y = 1 \\ y & if \quad x = 1 \\ 0 & otherwise \end{cases} \quad \text{(Drastic product)} \quad (1)$$

$$T_m(x,y) = \min(x,y) \quad \text{(Logical product)} \quad (2)$$

$$T_H(x,y) = \frac{xy}{x+y-xy} \quad \text{(Hamacher product)} \quad (3)$$

$$T_p(x,y) = x.y \quad \text{(Algebraic product)} \quad (4)$$

$$T_b(x,y) = \max(0, x+y-1) \quad \text{(Bounded product)} \quad (5)$$

$$T_E(x,y) = \frac{xy}{1+(1-x)(1-y)} \quad \text{(Einstein product)} \quad (6)$$

If T_1 and T_2 are two T-norms, then T_1 is weaker than T_2 (or T_2 is stronger than T_1) if for all $a,b\in[0,1]$, $T_1(a,b) \le T_2(a,b)$or simply $T_1 \le T_2$.

There are many papers on T-norms studying their properties. For example see [7, 9].

Despite being defined as binary functions, T-norms can be defined recursively, so that they also apply to larger numbers of clauses. Because of their associatively, for $a_1,\ldots a_n, a_{n+1} \in [0,1]$ the function can be defined as :

$T(a_1,\ldots a_n, a_{n+1}) = T(T(a_1,\ldots a_n), a_{n+1})$,

Definition 4.2: A function N: [0,1]→[0,1] is called a negation function if for all $x, y \in [0,1]$

($N1$) $N(1) = 0, N(0) = 1$ (Boundary conditions)

($N2$) $N(x) \leq N(y)$ if $x \geq y$ (monotonicty)

A special negation operator N that plays an important role in this framework is defined by $N(x) = 1 - x$

REFINEMENT BY T-NORM

Let w be a multi-meaning word with different meanings, say $w_1,\ldots,w_n$ that appears in a sentence together with the some connected words $a_1, a_2,\ldots,a_t$ from a large set of connected words described in section 2. Let $1 \leq j \leq n$ be given and for each $1 \leq i \leq t$, the connection degree of w_j to a_i be known as R_{ij}. To compute the connection degree of w_j to "*all*" a_i (i.e. R_j) , we model the conjunction operator "*and*" (see section 3) as usually used in fuzzy logic by T-norm T. Thus $R_j = T(R_{j1}, R_{j2}\ldots, R_{jt})$ for each $1 \leq j \leq n$. We notice that if the multi-meaning word w comes only with one connected word a_1 (i.e. t=1) in a sentence then $R_j = R_{j1}$ for each $1 \leq j \leq n$. Since we apply T-norms on R_{ij} s to compute each R_j, when the number of connected words is greater than two, R_j s would be small numbers close to zero and it would be difficult to compare them. To overcome the above situation and to compare them in a reasonable way, we "*normalize*" $R_1,\ldots,R_n$ in such a way that $R'_j = \frac{R_j}{\sum_{i=1}^{n} R_i}$ and clearly $\sum_{j=1}^{n} R'_j = 1$. Without loss of generality assume that $R_1 \leq R_2 \leq \ldots \leq R_n$ which implies that $R'_1 \leq R'_2 \leq \ldots \leq R'_n$.

As we already pointed out, based on this method w_n is chosen as the meaning of the multi-meaning word w. But here there is a degree of certainty of choosing w_n which depends on the distance between R'_n and the other $R'_j s$ This means that as the distance between R'_n and the other $R'_j s$ increases the meaning can be chosen with a greater degree of clarity. To measure the degree of clarity of chosen meaning, it is enough to consider the distance between the two largest R'_n and R'_{n-1} as the criterion of measurement of clarity degree. We set $d = R'_n - R'_{n-1}$. Clearly d∈ [0,1].Now to measure the degree of truth of the inferred meaning, we divide the unit interval [0,1] into some subintervals based on a logical procedure. These subintervals some how show the strongness or weakness of choosing the meaning of the multi-meaning word. We label these subintervals with linguistic terms in order to classify the degree of certainty. In this paper we divide the interval [0,1] to five subintervals and label them accordingly very weak, weak, more or less strong, strong, and very strong.

Now we choose two parameters a and b in [0,1] with negations N(a)=1-a and N(b)=1-b and divide the interval [0,1] into five subintervals as following: $[0,1] = [0,1-a) \cup [1-a,1-b) \cup [1-b,b)] \cup (b,a] \cup (a,1]$.

The choice of the parameter a may be content dependent, subjective, or a fixed number close to 1, for example $a = 0.85$. To use negation operator we find 1 - a = 0.15. Parameter b is chosen such that $[1-b,b]$ contains 0.5 and $2b - 1 > a - b > 1 - a$. If a=0.85 we may choose b=0.6 and the partition of [0,1] will be [0,1] = [0,0.15) ∪ [0.15,0.4) ∪ [0.4,0.6] ∪ (0.6,0.85] ∪ (0.85,1]

To choose the appropriate meaning of w, we argue as follow: compute $R'_n - R'_{n-1} = d$.

If $d \in (a, 1]$ we choose "*very strong*" the meaning of w as w_n.

If $d \in (b, a]$ we choose "*strong*" the meaning of w as w_n.

If $d \in [1 - b, b]$ we choose "*more or less strong*" the meaning of w as w_n.

If $d \in [1 - a, 1 - b)$ we choose "*weak*" the meaning of w as w_n.

If $d \in [0,1 - a)$ we choose "*very weak*" the meaning of w as w_n.

The number d measures the difference between two largest $R'_j s$. This shows how R'_n is compared to the others, but since we are not interested in the very small ones, we compute only $R'_n - R'_{n-1} = d \leq R'_n - R'_j$ for all j= 1,…, n – 2. Therefore d is a reasonable criterion to measure the farness of the data.

CASE STUDY

In this section we present two Persian sentences with same multi-meaning word and show how to apply T-norms and compare their result to get the correct meaning and its assigned labeled in each sentence.

A man in the forest survived by drinking a glass of **milk** *instead of water.*

A man in the forest helped a wild **lion** *to survive by drinking water.*

As we mentioned before, the word ([šir] " ") which appeared in the above Persian sentences is a multi-meaning word with correct meaning **milk** and **lion** respectively. We are about to find out its correct meaning by its connected under lined words showed in the above Persian sentences. The implemented system chose the multi- meaning word and its connected words and their weights for different meanings (R_{ij}) by

Table 1: R_{ij} s and R_j s

	Milk	Lion	Tap
(Glass)	R_{11}=0.78	R_{12}=0.02	R_{13}=0.43
(Drinking)	R_{21}=0.82	R_{22}=0.19	R_{23}=0.27
(Forest)	R_{31}=0.14	R_{32}=0.93	R_{33}=0.02
(Water)	R_{41}=0.40	R_{42}=0.30	R_{43}=0.85
T_w	0.0000	0.0000	0.0000
T_b	0.0000	0.0000	0.0000
T_h	0.1094	0.0176	0.0184
T_e	0.0166	0.0003	0.0006
T_m	0.1400	0.0200	0.0200
T_p	0.0358	0.0011	0.0020

Table 2: $R'_j s$

	Milk	Lion	Tap	d
N_h	0.7524	0.1211	0.1265	0.6259
N_e	0.9486	0.0171	0.0343	0.9143
N_m	0.7778	0.1111	0.1111	0.6667
N_p	0.9220	0.0282	0.0514	0.8706

Table 3: R_{ij} s and R_j s

	Milk	Lion	Tap
(wild)	R_{11} = 0.03	R_{12} = 0.94	R_{13} = 0.01
(Drinking)	R_{21} =0.82	R_{22} = 0.19	R_{23} = 0.27
(Forest)	R_{31} = 0.14	R_{32} = 0.93	R_{33} = 0.02
(Water)	R_{41} = 0.40	R_{42} = 0.30	R_{43} = 0.85
T_w	0.0000	0.0000	0.0000
T_b	0.0000	0.0000	0.0000
T_h	0.0243	0.1293	0.0066
T_e	0.0004	0.0282	0.0000
T_m	0.00300	0.1900	0.0100
T_p	0.0014	0.0498	0.0000

Table 4: $R'_j s$

	Milk	Lion	Tap	d
N_h	0.1517	0.8071	0.0412	0.6554
N_e	0.0140	0.9860	0.0000	0.9720
N_m	0.1304	0.8261	0.0435	0.6957
N_p	0.0273	0.9708	0.0019	0.9435

searching the database prepared before. The Table 1 shows the summarized results extracted by the system from database for the first Persian sentence and the results calculated manually by different T- norms on $R_{ij}s$. For example to compute R_1=0.1094 corresponding to Hamacher T-norm (T_h) we have:

$$\begin{aligned} T_h(0.78,0.82,0.14,0.40) &= T_h(T_h(0.78,0.82,0.14),0.40) \\ &= T_h(T_h(T_h(0.78,0.82),0.14),0.4) \\ &= T_h(T_h(0.67,0.14),0.40) \\ &= T_h(0.131,0.40) \\ &= 0.1094 \end{aligned}$$

Table 1 show T_w and T_b are not appropriate to use for this system. Other T-norms select the correct meaning milk according to the result in the corresponding column for multi-meaning word ([šir] " ") with different degree of clarity. To be more clear, we normalize the result of Table 1 (R_j s) to $R'_j s$ as mentioned before in section 5, for each T-norm T_h, T_e, T_m, T_p, now denoted by N_h, N_e, N_m, N_p respectively as shown in Table 2. In this table we also include the degrees of clarity d which is defined as the difference between two largest R'_j s

Results summarized in Table 1 and Table 2 show that the meaning of the multi-meaning word ([šir], " ") as **milk** has been selected correctly by all four T-norms, but with different degree of clarity which is shown in d-column. Based on our explanation in section 5, the degrees of clarity found by T_h and T_m are labeled "*strong*" and the degrees of clarity found by T_e and T_p are labeled "*very strong*". Since the degree of clarity found for Einstein Product T-norm (T_e) is greater than the other T-norms, we select T_e in our implemented system.

To evaluate the system more precisely we applied the system on the Persian sentence number 2 which has three common connected words with the Persian sentence number 1 and one different connected word. The results of this new sentence are summarized in Tables 3 and Table 4.

The results extracted from the second sentence are promising since not only the correct meaning **lion** has been selected, we see in Table 2 the weight of the correct meaning (milk) by N_e that is 0.9486 and incorrect meaning (lion) is 0.0171 and in Table 4 the weight of the correct meaning (lion) is 0.9860 and incorrect meaning (milk) is 0.0140. All these differences are because of only one uncommon connected word in both Persian sentences.

EVALUATION AND FUTURE RESEARCH

To support the idea described in this paper as discussed in section 2, a corpus containing 7800 sentences was prepared to extract connected words for chosen multi-meaning words. Considering different T-norms and applying them on the degrees of connection which were prepared before, leads us to the Einstein product T-norm (T_e) which is the best to refine the real meaning of a multi-meaning word in a sentence.

To evaluate the effectiveness of this T-norm (T_e) we applied it to over one hundred sentences and more than 95% of the results were correct. Positive point of this research is its text independency and the corpus mentioned above was prepared with this idea.

There are different ways to continue this research. In the first option we might consider more than one multi-meaning words instead of just one word, in a sentence. It could be more interesting if different meaning of the first multi-meaning word have different degree of connection for the second one.

The second option is considering a dynamic database for preparing CDCCW dynamically instead of the statically way which we have done so far.

REFERENCES

1. Ali, S.S., 1994. A logical Language for Natural Language Processing. Proc. of AI' 94, The Canadian Artificial Intelligence Conference, Alberta, Canada, pp: 16-20.
2. Bagga, A., J.Y. Chair and A.W. Biermann, 1997. The role of WordNet in the creation of a Trainable Message Understanding System, AAAI/IAAI, pp: 941-948.
3. Halliday, M., 1996. Spoken and Written Language, 2nd Edn, Oxford University Press.
4. Harabagiu, S.M., D.I. Moldovan and T. Yukawa, 1996. Testing Gricean Constraints on WordNet-based Coherence Evaluation System, AAAI.
5. Lang, J., 1990. Semantic Evaluation in Possibilistic Logic Application to Min-Max Discrete Optimization Problem.
6. Lauer, M., 1995. How much is enough: Data Requirements for statistical NLP, Proceedings of the second Conference of Pacific Association for computational Linguistics, Brisbane, Australia.
7. Mizumoto, M., 1989. Pictorial Representations of Fuzzy Connectives, part 1: Cases of T-norms, T-conorms and Averaging Operators, Fuzzy Sets and System, 31: 217-242.
8. Resnik, P., 1994. WordNet and Distributional Analysis, A Class-based Approch to Lexical Discovery, in Weir, C., S. Abney, R. Grishman and R. Weischedel (Eds.), Statically-based NLP Techniques, AAAI Press.
9. Schweizer, B. and A. Sklar, 1989. Probabilistic Metric Spaces, North-Holland, New York.
10. Smets, P. and P. Magrez, 1987. Implication in Fuzzy Logic, int j. of Approximate reasoning, pp: 337-347.
11. Ure, J., 1971. Lexical Density and Register Differentiation. In Perren, G. and J. Trim (Eds). Application of Linguistics, Cambridge University Press, London.