International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 2, March-April 2016 **ISSN 2278-6856**

Implementation of 6-W based precisiation structure for text summarization using V B .net

[1]PiyushPratapSingh, [2]Prof Jayashri Vajpai, [3]Prof V S Bansal (RETD.)

[1]Asst professor IT .CTS M G A H V Wardha MH

[2]Associate Professor Electrical Engineering Department Engineering College, Faculty of Engineering, JNVU Jodhpur Rajasthan

[3]Professor Electrical Engineering Department Engineering College, Faculty of Engineering, JNVU Jodhpur Rajasthan

Abstract

Text summarization is an emerging research domain of Computational Linguistics. It is an important area of research in human-machine interaction, which aims at interpretation of the natural language text by machine, at a level comparable with human being. This paper presents the design of a precisiation knowledge structure for text summarization using the concept of Six W's on V B .NET platform. This technique aims at developing a knowledge base which is compact and aids fast and efficient storage and search, because it overcomes the drawbacks associated with the key word and key phrase based searching employed by the approaches currently in prevalence. This approach classifies the natural language in a more refined and easy manner to access six attributes, viz. 6 W's, as opposed to the classification through parts of speech such as noun, verb etc. An illustrative example shows the efficacy of the new approach.

Key words NLP, CL AI, PNL

I Introduction

Natural Language Processing (NLP) in general and Computational Linguistics (CL) in particular, aims at interpretation of natural language text by computers at the level comparable with human being. Computers are nowadays used for preparation, acquisition, transmission, monitoring, storage, analysis and transformation of information. Hence, endowing them with the ability to understand and generate information expressed in natural language is a major field of research in Artificial Intelligence (AI). NLP is the fast developing research field due to its application in human- machine interaction (HMI) for communication with machines. It is essential to efficiently represent the natural language text in the HMI system so that information can be stored and retrieved easily and quickly. The most popular directions of research in computational linguistics are:

1. Morphological analysis of a variety of languages.
2. Grammar formalism and parsing programs.
3. Semantic extraction.
4. Development of specialized lexical resources.
5. Word sense disambiguation.
6. Automatic anaphora resolution.

Many theories and software have been developed and implemented recently for text summarization. This paper implements the six W concepts of Vajpai et al [9] using V B .net The design of this system precisiates the knowledge base (KB) of pre-stored sentences by employing six attributes. These six attributes are who, what, whom, when, where and why. This concept will be elaborated in section IV to develop a Structural Description Language (SDL). The paper also includes the state of art as understood from the study of literature. The details of implementation of SDL using VB.NET are presented in section V, followed by an illustrative example and conclusion.

II. State of art

The objective of computational linguistics is to develop models of language that can be implemented using computers. These models are used to develop application software for grammar correction, word sense disambiguation, compilation of dictionaries and corpora, intelligent information retrieval, automatic translation from one language to another, etc. Thus, CL is a major concern of both linguistics and computer science.

The most recent and powerful NLP software's working in public domain are Microsoft Word Grammar Checker by Microsoft Technologies, Google Translate by Google and Apple iSO 6 Siri by Apple. All of them deal with NLP, but their design and technology are different.

The process of MS Word Grammar Checker is based on simple pattern matching. The heart of the program includes an exhaustive list of phrases that are considered poor writing by many experts. The system identifies a list of possible correct phrases along with possible alternative wording for each phrase.

Google Translate generates a translation by searching for patterns in a large repository of documents to help decide on the best translation. By detecting pattern in documents that have already been translated by human translators, Google Translate selects an appropriate translation. This process of seeking pattern in large amount of text is called 'Statistical Machine Translation'.

Apple Siri is a speech recognition software that can process verbal command or instruction of user by employing a guided dialog to domain and task model, which is also connected to the web services and application program interface (API). This facilitates speech based command following in natural language. However, this has achieved limited success.

Information retrieval (IR), Question answering (QA) and text summarization (TS) are the key application areas of computational linguistics. The following researchers have proposed the most important design initiatives and developed theories in these areas of research.

Zadeh[1-2] has proposed a breakthrough technology viz., fuzzy logics which has influenced computational linguistics. This is called Precisiated Natural Language (PNL) which is completely different from the conventional concept. PNL tries to summarize the given text or to classify the language. Zadeh [3-4] further proposed the concept of fuzzy or generalized constraints and integrated it with the PNL to develop Generalised Constraints Language (GCL). The concept of protoform analysis was used to develop the structure of PNL. However, it is yet to be implemented in natural language computing due to the absence of precise knowledge structure to implement fuzzy constraints. Zadeh also proposed PNL based QA system to overcome the problem of searching world knowledge and gave the concept of calculation of relevance by using fuzzy logics.[5]

Thint et al [6-7] have implemented the PNL in Question answering system. Thint et al started their research by implementing semi automated detection of PNL protoforms by using its rules and conventional If-then-else rules and later advanced to an information retrieval system using PNL based reasoning and deduction based reasoning. In this system the search was based on Key phrase and complex rules were derived, which made the system complicated to implement as different phrases require different set of rules. Thint et al then restricted their domain of application to the QA system and classification of natural language was not done.

Jana et al [8] have proposed a QA system in which many precise matching prepositions can be deduced from the known propositions. In this approach also, the implementation of rules was complicated and classification of natural language could not be carried out.

Recently, Ledeneva *et al*. [12-15] have successfully employed the word sequences from the given text for selecting the text fragments for composing the summary. Initially Ledeneva *et al*. suggested a typical automatic extractive summarization approach composed of term selection, term weighting, sentence weighting and sentence selection. However, their approach is not able to classify the natural language texts that are contained as sub-sequences in maximal frequent word sequences.

Garcia *et al*. [16-18] have extracted the sequences of words from the language text by using the conventional key based search, which is exhaustive in nature. In this work, sentences were extracted using unstructured learning approach without developing a KB, thus eliminating the possibility of experience based learning.

Vajpai et al [9-10] have proposed a 6 W's based concept that overcomes many drawbacks of the above approaches. This classifies the entire NL text in six categories, facilitating the summarization of text and avoiding key word base search. The implementation and further details of this approach are discussed in the next section.

III. 6 – W Based Structural Description Language

Rudyard Kipling, an indefatigable globe trotter, discovered that all the knowledge of global activities can be described by a small group of information carriers. The six principal elements of this group are [8]

Who – The actors of activities

What- The actions carried out by the actors

Whom- The object to which the activities are directed

When- The time of activities

Where- The place of activities

Why- The reasons which prompted the occurrence of activities

These serve as the primary elements of description in the proposed Structural Description Language (SDL) as shown in Fig. 1. They can be used for effective summarization of text with minimal loss of information. The implementation of the proposed SDL by usingVB.NET platform is discussed in the next section.

IV. Implementation of SDL usingVB.NET

VB .net is a powerful tool for developing software and has been used in this research for the design of 6W based Text Summarization Software because of its friendly graphical user interface and strong database connectivity. An important reason for the selection ofVB.NET is that it is compatible with Windows because it also belongs to the Microsoft Inc. It is notable that around 90 % people of the world use Window's on their laptops and desktops [11].VB.NET is easy to implement in the Windows environment.

The main aim of the designed software is to classify text to the six attributes or six W's for text summarization as shown in .Fig 1.

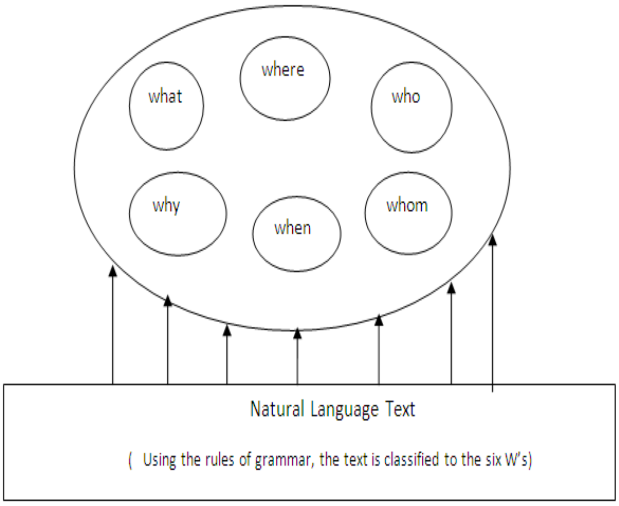

where
what
who
why
when
whom
Natural Language Text
(Using the rules of grammar, the text is classified to the six W's)

Fig 1 Elements of Structural description language

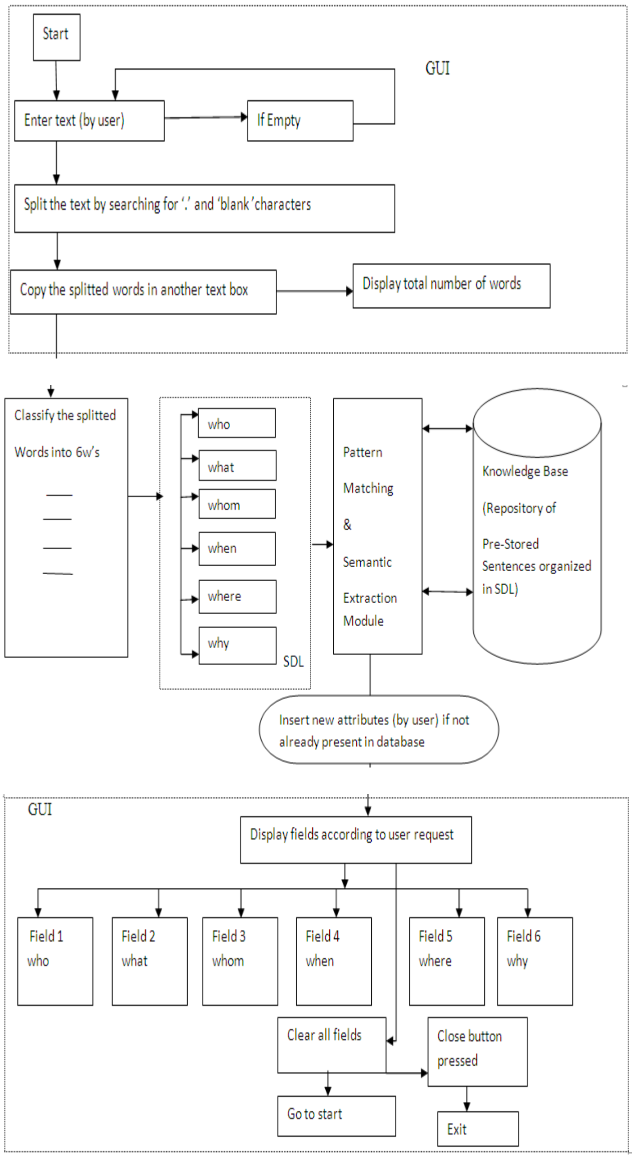

Start
GUI
Enter text (by user)
If Empty
Split the text by searching for '.' and 'blank' characters
Copy the splitted words in another text box
Display total number of words
Classify the splitted Words into 6w's
who
what
whom
when
where
why
SDL
Pattern Matching & Semantic Extraction Module
Knowledge Base (Repository of Pre-Stored Sentences organized in SDL)
Insert new attributes (by user) if not already present in database
GUI
Display fields according to user request
Field 1 who
Field 2 what
Field 3 whom
Field 4 when
Field 5 where
Field 6 why
Clear all fields
Close button pressed
Go to start
Exit

Fig 2. Data flow diagram

The functioning of the designed software is depicted in the Data flow diagram shown in fig 2. The DFD comprises of four main sub modules:

- The Knowledge Base (KB)
- Pattern matching and semantic extraction module
- The SDL module
- The graphical user interface (GUI).

KB is the repository of pre-stored written words and sentences classified in the SDL. The pattern matching and semantic extraction module classifies the written text into the six fields. The GUI of the software helps in entering text, splitting sentences into six Ws and displaying the result.

The major steps in its implementation are as follows:

1. 1 The initial window of the software asks the user to enter text in the text box.
2. After entering the text user clicks on the split button shown in fig 3.
3. The software then splits the text into words by searching for '.' & 'blank' character and arranges the different words in an array.
4. The software displays the splitted words in a display box for further search and also displays the total count of words.
5. The classification is carried out for these splitted words by matching them with the pre-stored words or group of words in the knowledge base with respect to six W structure. Semantic extraction is applied on the classified words by pattern matching with a pre-stored knowledge repository of sentences described in the SDL comprising of the six W's organized as follows :

Who- generally the first noun or name.

What – verb (action)

Whom –second noun (name)

When –time.

Where – place.

Why – reason for action

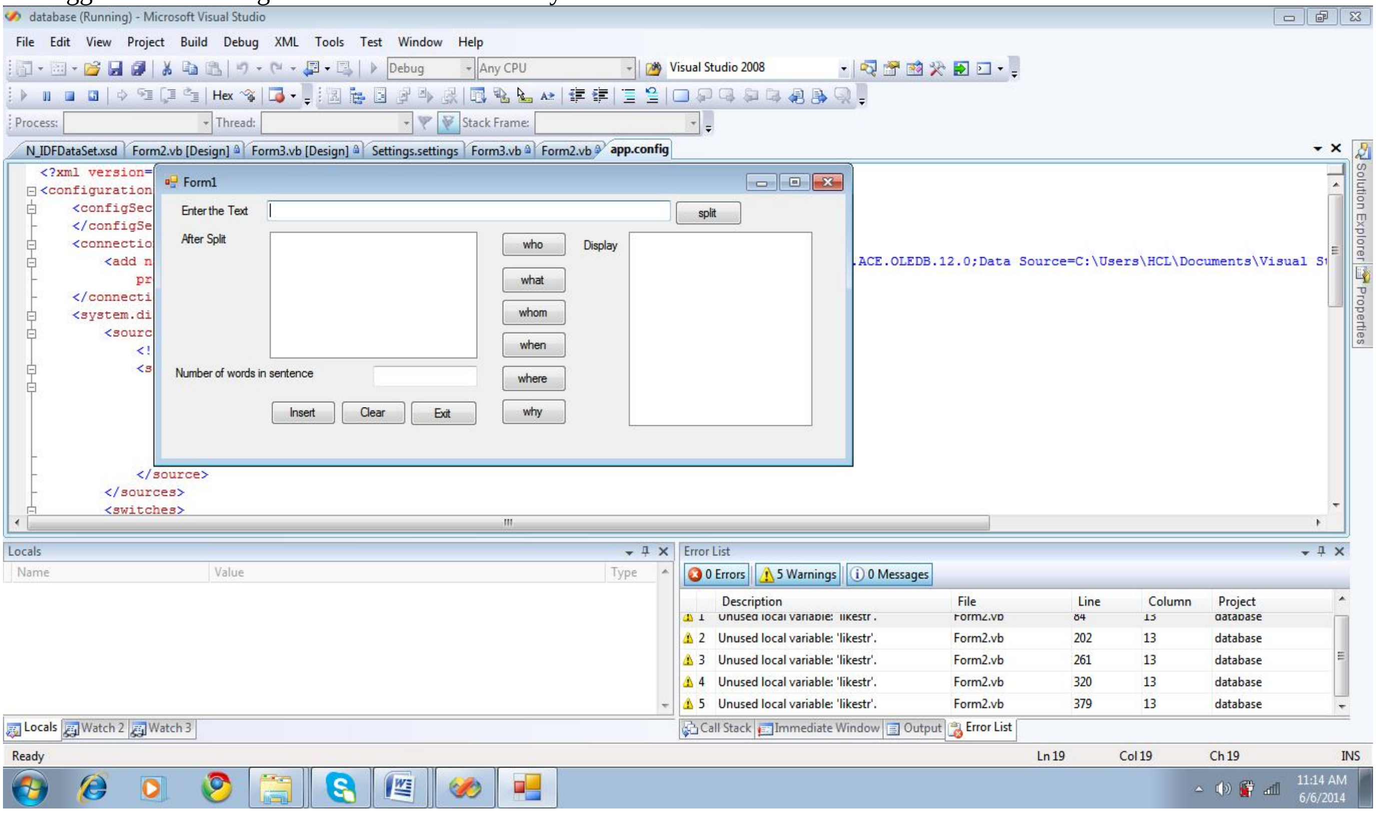
This is the most important step which classifies the entire sentence in the 6 fields as shown in fig 4

6. The software extracts the words or group to the respective W attributes and displays them in related fields of the result window shown in fig 5.

7. The user can update the KB as per requirement by using insert button shown in fig 6, and directly add new

6W attributes, which helps in continuously updating the Knowledge base, thus progressively enhancing its content.

The success of the software also depends on this step, i.e. the bigger the knowledge base better the efficiency of the software. There are three more buttons in the software. The ‘Clear’ button- clears the entire window and resets the process, the ‘Exit’ button- exits the software and the ‘Close’ button closes the insertion window.



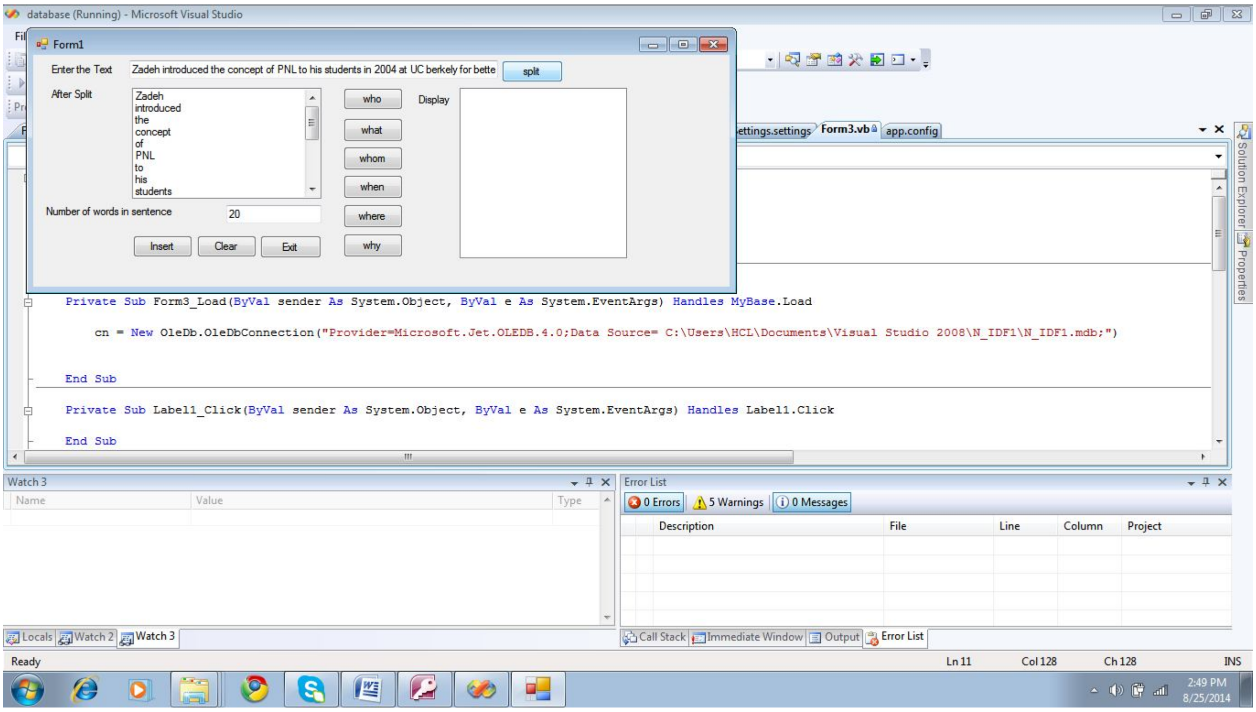
```
<?xml version=
<configuration
    <configSec
    </configSe
    <connectio
        <add n
            pr                                        .ACE.OLEDB.12.0;Data Source=C:\Users\HCL\Documents\Visual St
    </connecti
    <system.di
        <sourc
            <!
            <s

            </source>
        </sources>
        <switches>
```

Error List

0 Errors | 5 Warnings | 0 Messages

	Description	File	Line	Column	Project
1	Unused local variable: 'likestr'.	Form2.vb	84	13	database
2	Unused local variable: 'likestr'.	Form2.vb	202	13	database
3	Unused local variable: 'likestr'.	Form2.vb	261	13	database
4	Unused local variable: 'likestr'.	Form2.vb	320	13	database
5	Unused local variable: 'likestr'.	Form2.vb	379	13	database

Fig 3 Initial window



Enter the Text: Zadeh introduced the concept of PNL to his students in 2004 at UC berkely for bette

After Split:
Zadeh
introduced
the
concept
of
PNL
to
his
students

Number of words in sentence: 20

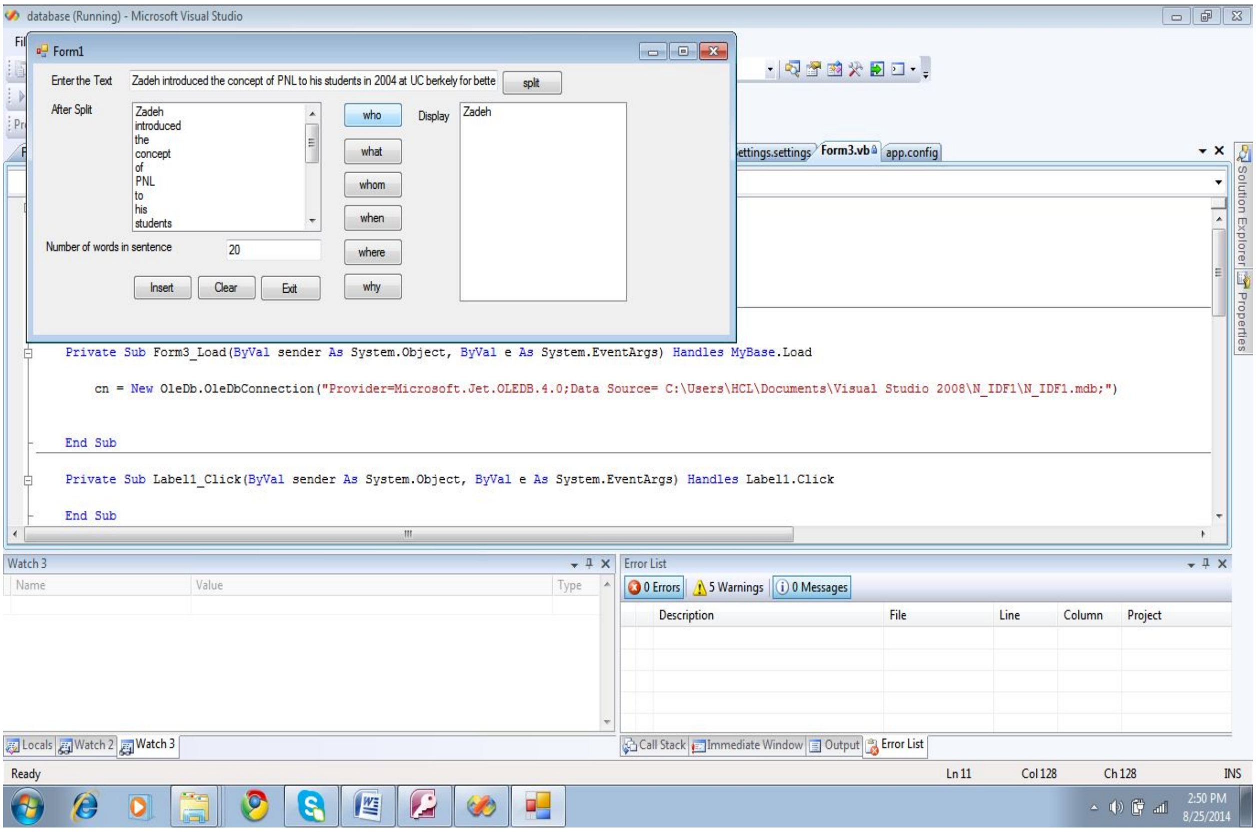
```
Private Sub Form3_Load(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles MyBase.Load

    cn = New OleDb.OleDbConnection("Provider=Microsoft.Jet.OLEDB.4.0;Data Source= C:\Users\HCL\Documents\Visual Studio 2008\N_IDF1\N_IDF1.mdb;")


End Sub

Private Sub Label1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Label1.Click

End Sub
```

Error List

0 Errors | 5 Warnings | 0 Messages

Description	File	Line	Column	Project

Fig 4 Splitting window



Enter the Text: Zadeh introduced the concept of PNL to his students in 2004 at UC berkely for bette

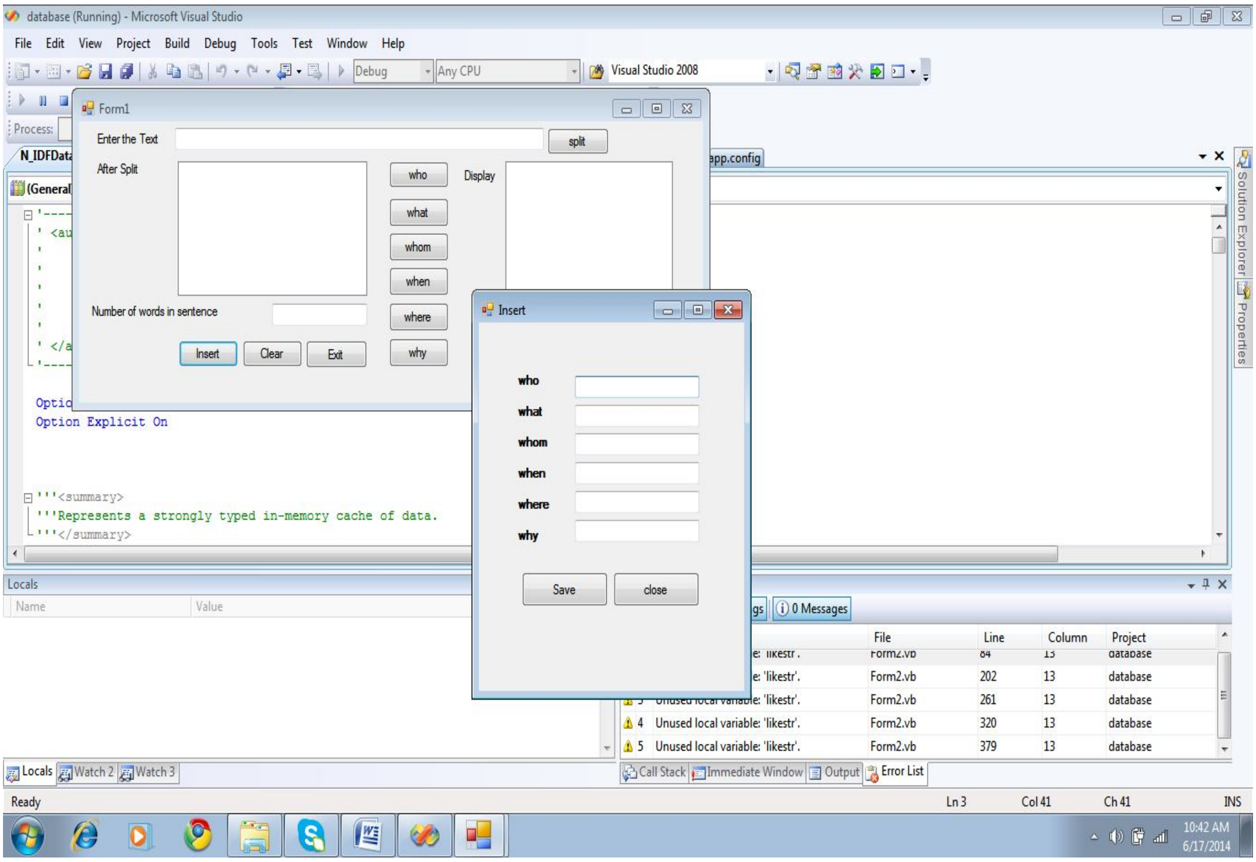
After Split: Zadeh introduced the concept of PNL to his students

Display: Zadeh

Number of words in sentence: 20

```
Private Sub Form3_Load(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles MyBase.Load

    cn = New OleDb.OleDbConnection("Provider=Microsoft.Jet.OLEDB.4.0;Data Source= C:\Users\HCL\Documents\Visual Studio 2008\N_IDF1\N_IDF1.mdb;")

End Sub

Private Sub Label1_Click(ByVal sender As System.Object, ByVal e As System.EventArgs) Handles Label1.Click

End Sub
```

Fig 5 Result window



Insert: who, what, whom, when, where, why

```
Option Explicit On

'''<summary>
'''Represents a strongly typed in-memory cache of data.
'''</summary>
```

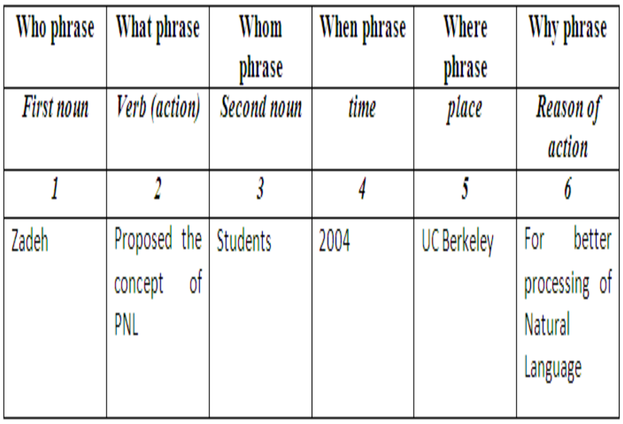
	Description	File	Line	Column	Project
4	Unused local variable: 'likestr'.	Form2.vb	320	13	database
5	Unused local variable: 'likestr'.	Form2.vb	379	13	database

Fig 6. Insertion window

V. Illustrative Example

Let us consider the following natural language sentence in English language “Zadeh proposed the concept of PNL to his students in 2004 at UC Berkeley for better processing of Natural Language”. Now, using 6-W based SDL on the sentence the classification depicted in the table 1 is obtained.

Table 1 Classification of the example text



Who phrase	What phrase	Whom phrase	When phrase	Where phrase	Why phrase
First noun	*Verb (action)*	*Second noun*	*time*	*place*	*Reason of action*
1	*2*	*3*	*4*	*5*	*6*
Zadeh	Proposed the concept of PNL	Students	2004	UC Berkeley	For better processing of Natural Language

The sentence is updated in the knowledge base in the SDL format. This format of sentence provides quick answer to any query about the activity described in the sentence. For example, let the query be “Who proposed PNL?” By clicking the Who button on the software window, the answer to this query can readily be found in the element corresponding to Who in the SDL.

VI Conclusion

The proposed SDL based software providesVB.NET based implementation of the concept of 6 W’s based text precisiation technique. It classifies the natural language into six attributes for classification of language components into primary elements of information, resulting in more efficient representation, searching and question answering. This software and technique can be used for text summarization, information retrieval and automated question answering. Unlike other prevalent techniques, this theory does not use key word or phrase for searching. Work is in progress to further enhance the capability of this software by incorporating a semantic net based detailed structure for further elaboration of the information of the individual W’s.

VI. Appendix

Microsoft first started Visual Basic in early 1990s and the project name was "Thunder". After the launch of VB 5.0, it crosses all the boundaries and won the best RAD Tool award by beating PowerBuilder in 1998. VB 5.0 came out with some great enhancements but definitely this time VB.Net has come with revolutionary changes to make it suitable for next generation of application development in 2002.

Problems with VB 6.0

1. No capabilities for multithreading.
2. Lack of implementation inheritance and other object oriented features.
3. Poor error handling capabilities.
4. Poor integration with other languages such as C++.
5. No effective user interface for Internet based applications.

Principal features of VB.NET

1. Full support for object oriented programming.
2. Structured error handling capabilities.
3. Access to .NET Framework.
4. Powerful unified Integrated Development Environment (IDE).
5. Inherent support for XML & Web Services.
6. Better windows applications with Windows Forms.
7. New Console capabilities of VB.NET.
8. New Web capabilities with Web Forms.
9. Immense power of tools & controls (including Server Controls).
10. Interoperability with other .NET compiled languages.
11. Better database programming approach with ADO.NET.

VB.NET is implemented by Microsoft's .NET framework. Therefore, it has full access to all the libraries in the .Net Framework. It's also possible to run VB.NET programs on Mono, the open-source alternative to .NET, not only under Windows, but even Linux or Mac OSX.

References

[1]. L. A. Zadeh, “From computing with numbers to computing with words from manipulation of measurements to manipulation of perceptions”, International Journal for Applied Math & Computer Science., Vol. 12/3: pp. 307-324, 2001.

[2]. L. A. Zadeh, “A new direction in AI - toward a computational theory of perceptions”, A.I. Magazine, Spring 2001.

[3]. L. A. Zadeh, “Precisiated natural language,” AI Magazine, 25(3), 2004, pp. 74-91.

[4]. L. A. Zadeh, “Toward a generalized theory of uncertainty (GTU)” -an outline in Information Sciences, 172, 2005, pp. 1-40.

[5]. L. A. Zadeh “From Search Engines to Question-Answering Systems: The Problems of World knowledge Relevance, Deduction, and Precisiation” * http://www.springer.com/978-3-540-34780-4 2006

[6]. M. Thint, M S Beg, Z. Qin, “PNL-enhanced Restricted Domain Question Answering System,” IEEE International Conference on Fuzzy Systems, London, UK, July 2007.

[7]. M Thint, M. S. Beg, Z. Qin, M. “Deduction Engine Design for PNL-based Question Answering System,” World Congress of the International Fuzzy Systems Association , 2007.

[8]. J. shafi ,A. ali “Defining relations in precisiation of natural language processing for semantic web” IJCSE ISSN:0975-3397 Vol .4 no 05 May 2012.pp 72

[9]. K.R. Chowdhary, J. Vajpai, V.S. Bansal, “Natural language text compression using 5-W based presiciation structure” ,ECTN-12, 24 25 March 2012.

[10]. J. Vajpai, V.S. Bansal, P.P Singh, “Computer assisted multilingual translation for global communication”, National Conference on “New Advances in Programming languages and their implementations”, March 15-16, 2013

[11]. “Usage of windows operating system in public domain” http://wikipedia.org/wiki/Usage_share_of_operating_systems 2013

[12]. Y. Ledeneva, “Recent Advances in Computational linguistics”, Mexico Informatica 34 2010

[13]. Y. Ledeneva, “Effect of Preprocessing on Extractive Summarization with Maximal Frequent Sequences”, MICAI-08, LNAI 5317,pp. 123-132, Mexico, Springer-Verlag, ISSN 0302-9743, 2008.

[14]. Y. Ledeneva, A. Gelbukh, R. G. Hernández, “ Terms Derived from Frequent Sequences for Extractive Text Summarization”, CICLing-08, LNCS 4919, pp 593-604, Israel, Springer-Verlag, ISSN 0302-9743, 2008.

[15]. Y. Ledeneva, A. Gelbukh, R. G. Hernández, “Keeping Maximal Frequent Sequences Facilitates Extractive Summarization”, In: G. Sidorov et al (Eds). CORE-2008, Research in Computing Science, vol. 34, pp.163-174, ISSN 1870-4069, 2008.

[16]. A. Gelbukh and G. Sidorov “Approach to construction of automatic morphological analysis systems for inflective languages with little effort”, Lecture Notes in Computer Science, N 2588, 2003, ISSN 0302-9743, Springer-Verlag, pp. 215–220

[17]. A Gelbukh ,G Sidorov., S Y Han, “Evolutionary Approach to Natural Language Word Sense Disambiguation through Global Coherence Optimization”, WSEAS Transactions on Communications, ISSN 1109-2742, Issue 1 Vol. 2, pp. 11–19, 2003.

[18]. A Gelbukh., I Bolshakov, “Internet, a true friend of translator”, International Journal of Translation, ISSN 0970-9819, Vol. 15, No. 2, pp. 31–50, 2003.

[19]. K. Toutanova, D. Klein, C. Manning, and Y. Singer, “Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network,” in Proc. of HLT-NAACL 2003, pp. 252-259.

[20]. “Microsoft Word grammar checker” http://office.microsoft.com/en-in/word-help/check-spelling-and-grammar-P010117963.aspx 2011

[21]. “Google Translate” http://translate.google.com. 2014

[22]. “Apple siri software for iSO phones” http://www.zdnet.com/ applesiri 2013