An Approach to Detect Slang Words in e-Data

Alok Ranjan Pal
Dept. of Computer Science and Engineering
College of Engineering and Management, Kolaghat
West Bengal, India
chhaandasik@gmail.com

Diganta Saha
Dept. of Computer Science and Engineering
Jadavpur University
Kolkata, India
neruda0101@yahoo.com

***Abstract*—The proposed approach deals with the detection of slang words in electronic data in different communication mediums like internet, mobile services etc. But in the real life, the slang words are not used in complete word forms always. Most of the times, those words are used in different abbreviated forms like sounds alike forms or taboo morphemes. This proposed approach detects those abbreviated forms also using semi supervised learning methodology. This learning methodology derives the probability of a suspicious word to be a slang word by the synset and concept analysis of the text.**

Keywords- Natural Language Processing (NLP); Slang word; Suspicious word; Synset; Concept

I. INTRODUCTION

World Wide Web and Telecommunication system play a vital role in this fast and modern era. One of the major activities is information sharing. People can communicate with other via e-mail, chatting, community forums, SMS etc. By a single click or by pressing a single button on our mobile phones, we can communicate with people, who are far away from us. In the field of study, research, business, sports, entertainment, national and international affairs etc. these electronic mediums have made a radical change.

But these facilities have some negative influences also on the society. At the time of information sharing, sometimes people use abusive words. In different public community forums, these slang words openly appear. This type of malfunctioning makes the web polluted. Different communities use this medium to spread rumor and violence. Recently, Governments of different countries are taking different steps to protect the malfunctioning over Internet and Telecommunication system. As some effective measures, use of few words is banned in SMS, selected sites are blocked and discussion on few topics on community forum etc. are prohibited in different countries.

This algorithm would detect the slang words, used in different texts at the time of its submission onto the web or any network.

Organization of rest of the paper is as follows: Section 2 is about motivation of our paper; Section 3 describes the background; Section 4 depicts the proposed approach in detail; Section 5 depicts experimental results; Section 6 represents conclusion of the paper.

II. MOTIVATION

Recently, different effective measures have been taken by the Governments of different countries around the world. Pakistani government has banned few web sites, implied restriction on use of near about 1700 words in SMS. USA and CHINA also have taken different steps to resist the malfunctioning. Most recently, Government of India warns few community sites to become more responsible about their social consciousness. They should monitor the data, which are handled by their system.

This proposed algorithm would detect a slang word in a text, if this is used as a script in the specific method, through which the data is passed.

III. THEORITICAL BACKGROUND

Natural Language Processing (NLP) [1] play an important role in different real life applications. So many research works are carried out in different fields of NLP, as Information Retrieval (IR), Automated Classification [7], Language Translation by Machine [4, 5, 6], Word Sense Disambiguation [9, 10, 17], Part of speech Tagging [15, 16], Anaphora Resolution [11], Paraphrasing [12], Malapropism [13], Collocation Testing [14] etc. These research works greatly depend on some knowledge driven methods [8]. WordNet [2, 18, 20] is a machine readable dictionary, used as a strong knowledge base now a day. In this dictionary words are arranged semantically instead of alphabetically. Words of symmetric sense are grouped together into a set, called synset and each synset represents a distinct sense, called concept [19].This organisation of words play a vital role in different research areas of NLP like Automatic Summarisation, E-Learning [3], Automatic Medical Diagnosis etc.

Our approach adopts the idea of resolution of a sense from a given text to detect a slang word in that text.

IV. PROPOSED APPROACH

This approach handles the slang words in four ways. First, each word of the input text is compared with the entries of a slang word Database. This Database is initially populated with some usual slang words. If any word of the input text is completely matched with any entry of the Database, the process stops proceeding (Figure 10) with a message.

An additional job is performed here. The sense of the discussion (concept) [19] is derived from the text.

ISBN: 978-81-923777-9-7

Secondly, each word of the input text is compared with the entries of another Database, contains a few commonly used slang words and the words, which sound like those slang words (Figure 11). If the input text contains any sounds-alike slang word, the algorithm detects that word and stops proceeding.

Thirdly, the algorithm handles the suspicious words using sliding window mechanism (Figure 5).

Finally, the learning ability of the algorithm is handled by the derived concept and the suspicious word Database (Figure 13).

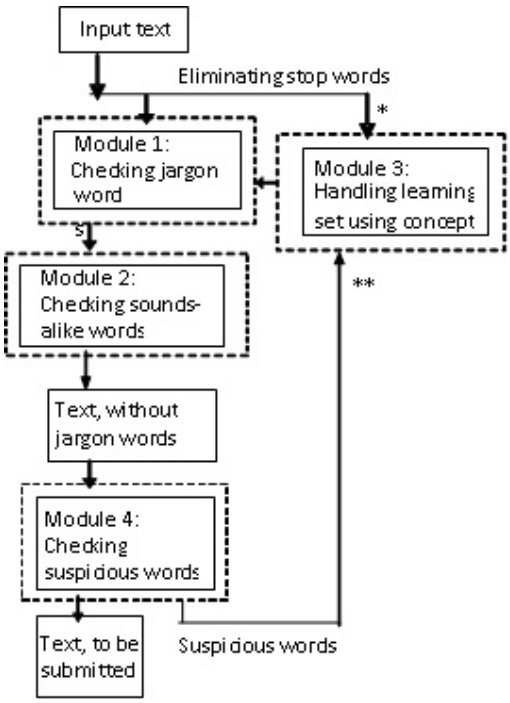


Figure 1. Block diagram of the overall procedure.

Algorithm 1: This algorithm takes a text as input and generates a text, free from slang word. It performs four jobs in four different modules (Figure 1). In module 1, the slang words are detected. In module 2, the sounds-alike slang words are detected. The suspicious words are detected in Module 4 and the learning set is enriched in Module 3.

Input: Input text.

Output: Text to be submitted.

1. Stop words are eliminated from the input text.
2. Text, with only meaningful words, is created.
3. This text is passed to
   a. Module 1 for detecting the slang words.
   b. Module 3 for deriving the concept from the text.
4. Text, without completely matched slang words, is obtained from Module 1.
5. The text from Module 1 is passed to Module 2, where the sounds-alike slang words are checked.
6. The text from Module 2 is passed to Module 4, where the occurrence of any suspicious word is checked.

   If any suspicious word is found in the text, it is passed to Module 3, where the learning set is enriched.
7. Text, without any slang or suspicious word is derived for further use.
8. Stop.

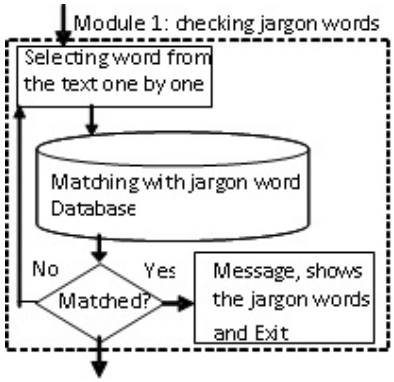


Figure 2. Block diagram of the slang word detection module.

Module 1: Algorithm 2: This algorithm checks each word of the text with the entries of a Database of slang words. If any word is completely matched with any entry (Figure 2), this algorithm shows that word and stops proceeding. The Time Complexity of the algorithm is $O(n^2)$. Complexity of picking up the n number of words from the text is of O(n) and checking the each word with the entries of the Database is O(n).

Input: Text, containing only meaningful words.

Output: Text, without slang words.

1. Repeat steps 2, 3 and 4 for each word of the text.
2. A word from the text is taken.
3. The word is matched with each entry of the slang word Database.
4. If the word is matched with any entry, the algorithm stops proceeding and Exit.

   Else: Goto step1 loop.
5. Stop.

ISBN: 978-81-923777-9-7

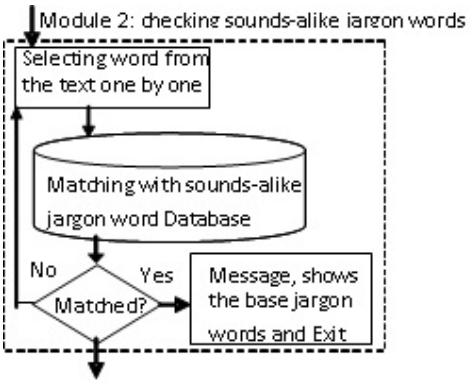


Figure 3. Block diagram of the sounds-alike slang word detection module.

Module 2: Algorithm 3: This algorithm checks each word of the text with the entries of a sounds-alike slang words Database (Figure 3). If any sounds-alike word is matched with any entry, that is treated as a slang word and the algorithm stops proceedings. The Time Complexity of the algorithm is $O(n^2)$, derived in the same way as in Module 1: Algorithm 2.

Input: Text, free from completely matched slang words.

Output: Text, without slang words.

1. Repeat steps 2, 3 and 4 for each word of the text.
2. A word is taken from the text.
3. The word is matched with each entry of the sounds-alike slang word Database.
4. If the word is matched with any entry,

   A message displays the actual slang word, which must be changed to proceed and Exit.

   Else: Goto step 1 loop.
5. Stop.

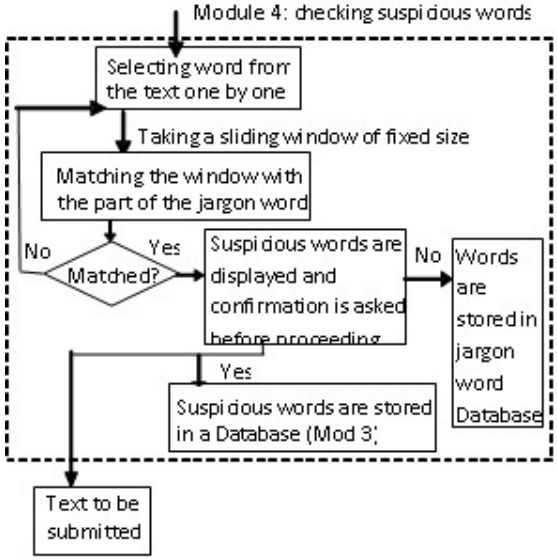


Figure 4. Block diagram of the suspicious word detection module.

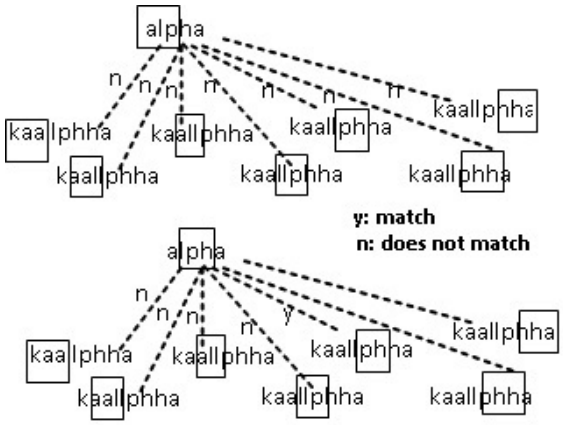


Figure 5. Block diagram of the sliding window mechanism.

Module 4: Algorithm 5: This algorithm checks the suspicious words (Figure 4) from the text, using a sliding window of some fixed length. If any word from the text is matched partially with any part of a slang word, that word is treated as a suspicious word. The Time Complexity of the algorithm is $O(n^2)$ due to the nesting of operations at Step 1 and Step 4.

Input: Text, without slang word.

Output: a) Text, which can be used for proceeding and

b) List of Suspicious words.

1. Repeat step 2, 3 and 4 for each word of the input text.
2. A word is taken.
3. A character window (Figure 5) of some fixed length is taken and set at the beginning of the word.
4. Repeat step 5 till the window lies within the word length.
5. If the window is matched with any part of the slang words, a message, mentioning the suspicious word is displayed and waits for confirmation to proceed.

   If the user confirms the suspicious word as a slang word, the word is stored in slang word Database directly and exit.

   Else, The suspicious word is passed to module 3 for enriching the learning set.

   Else,

   The window is shifted right by one character.
6. The text is displayed.
7. Stop.

ISBN: 978-81-923777-9-7

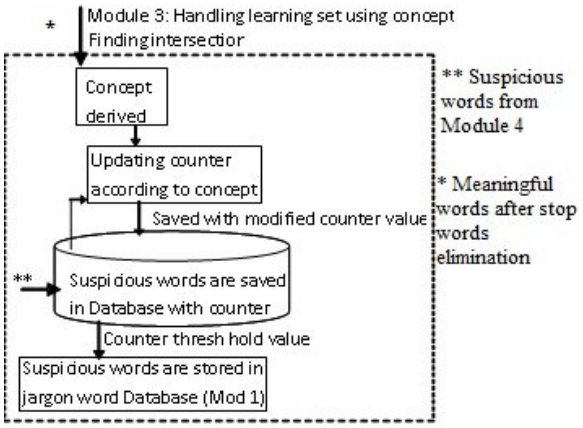


Figure 6. Block diagram of the learning mechanism.

Module 3: Algorithm 4: This algorithm handles the learning set, which increases the efficiency of detection of slang words. The maximum Time Complexity of the algorithm is $O(n^2)$, which is evaluated at step 1.

Input: a) Text, containing the only meaningful words from the input text and

b) Suspicious words from module 4.

Output: Slang words would be stored in slang word Database (Module 1).

1. Intersection (Figure 6) is performed between the words of the input text and the different synsets.
2. The concept is derived from the highest value of intersection.
3. Some predefined weight is assigned to the derived concept.
4. Suspicious words from Module 4 are stored in a Database with some counter.
5. Repeat steps 6, 7 and 8 for each suspicious word.
6. The counter value of a suspicious word is taken.
7. The counter value is updated by the weight, assigned for the concept.

   If the counter value of a suspicious word crosses some thresh hold value, the suspicious word is treated as a slang word and is stored in slang word Database (Module 1) for further decision making.

8. Stop.

V. Output and Discussion

The experiment was started with a table of common slang words (Figure 7) and a table, consisting of different synsets and the related concepts (Figure 8). Each concept carries some weight ("assgval" in Figure 8), which is used for learning. For experiment few Greek words are considered as slang words.

JID	Slng
10	alpha
11	beta
12	gamma
13	delta
21	**********
37	epsilon
40	lambda
41	upsilon
iber)	

Figure 7. The table, containing few slang words.

cid	concept	synset	assgval
1	Movie	$song$actor$actress$director$film$cam	10
2	Sports	$match$cricket$football$player$ground$	7
3	Business	$import$export$sell$purchase$shop$ma	6
4	Education	$teacher$student$subject$class$vacati	3
ber)			

Figure 8. The table of different synsets and related concepts.

In Figure 9, it is depicted; the input text is displayed as the output text as there is no slang word in the input text.

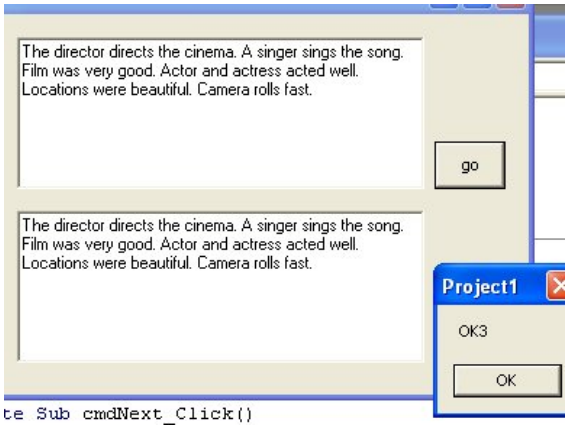
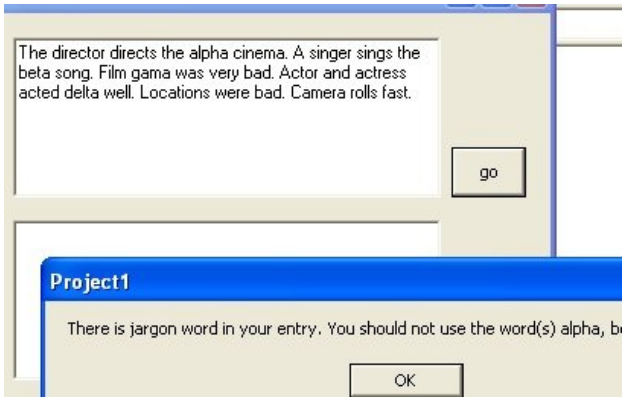


Figure 9. The input text is submitted as it does not contain any slang word.

In Figure 10, it is depicted; the input text contains few slang words.



ISBN: 978-81-923777-9-7

Figure 10. Few slang words are detected in the input text. So, it is not submitted as output text.

In Figure 11, it is depicted; the input text contains few words, which sound like slang words.

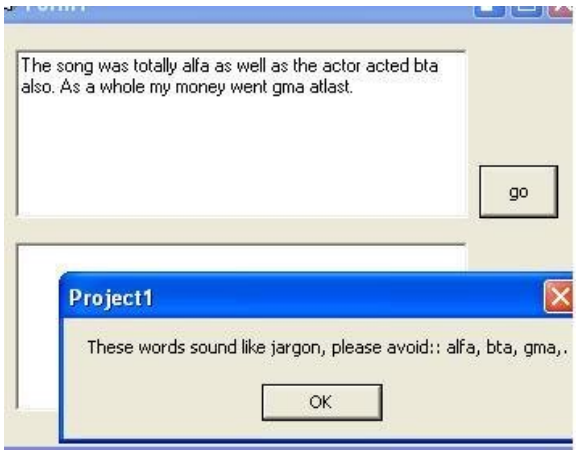


Figure 11. Few sounds-alike slang words are detected in the input text.

In Figure 12, it is depicted; the input text contains few suspicious words, which do not match with the stored slang words completely but partially.

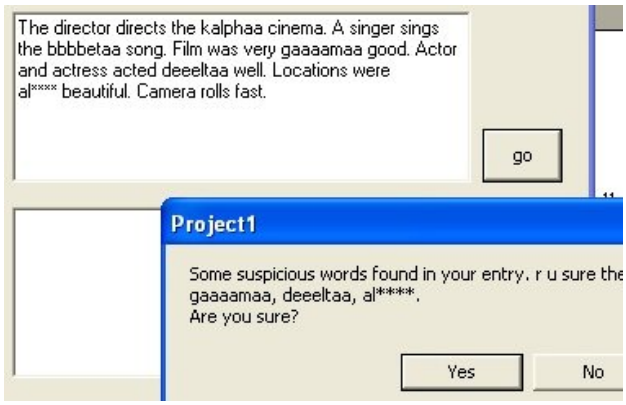
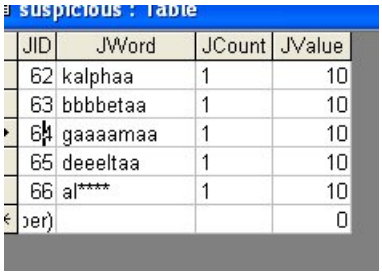


Figure 12. A message shows the suspicious words and waits for confirmation to proceed.

When a text, containing such slang words(Figure 12) is submitted, the suspicious words are detected and alerts the user for verification.

If the user does not want to proceed with those suspicious words, the process will be stopped and the derived suspicious words would be treated as slang words without any further verification and would stored in the slang word Database for further decision making. Otherwise, the suspicious words would be stored in a Database with some counter ("JCount" in Figure 13). Counter denotes the frequency of occurrence of that suspicious word in different contexts.



JID	JWord	JCount	JValue
62	kalphaa	1	10
63	bbbbetaa	1	10
64	gaaaamaa	1	10
65	deeeltaa	1	10
66	al****	1	10
ber)			0

Figure 13. Suspicious words are stored with counter value, shown in “JCount” column.

Now, the learning ability of the algorithm is discussed with example.

Figure 8 depicts the different synsets and associated concepts which are manually tagged for experiment. From real life scenario, it is obvious that according to the context of the input text, the probability of a suspicious word, to be a slang word, is changed. In the “assgval” column, these different values are stored from real life knowledge.

In figure 12, the concept of the input text is “Movie”. So, the suspicious words were stored in to the suspicious-word Database (Figure 13), with pre-assigned weight 10 (column "JValue" in Figure 13).

Another text is considered about “Sports” and the suspicious words are stored in suspicious word Database with pre-assigned weight 7 (Figure 14).

If any suspicious word is used again and again in different contexts, the associated "JValue" would be increased (Figure 14). If the "JValue" of any particular suspicious word crosses some threshold value, that suspicious word would be treated as slang word and that would be stored in the slang word Database (Figure 15). Experimentally, threshold value is taken 50 here.

JID	JWord	JCount	JValue
71	eeeeepsilon	1	7
72	theeeeta	1	7
73	lambddddaaa	1	7
74	uuupppsilon	1	7
79	kalphaa	4	40
80	bbbbetaa	4	40
81	deeeltaa	4	40
82	al****	4	40
ber)			0

Figure 14. “JValue” of a suspicious word is increased, as it is used repeatedly.

ISBN: 978-81-923777-9-7

	Slng
10	alpha
11	beta
12	gamma
13	delta
21	**********
37	epsilon
40	lambda
41	upsilon
59	kalphaa
60	bbbbetaa
61	deeeltaa
62	al****
imber)	

Figure 15. Suspicious word is treated as slang word, as its 'JValue' crosses threshold.

In this way, as the algorithm would be trained by different texts of different contexts, the slang word Database would be stronger. The "assgval" is decided from the real life scenario according to the different concepts (in Figure 8). All the user assigned values might vary situation wise.

VI. CONCLUSION AND FUTURE WORK

This algorithm detects the slang words, as well as the suspicious words from a text, which is used for conversation in any open medium. If the learning mechanism is handled with proper synset and probability analysis, the algorithm would solve a big problem of recent day.

But, in some cases like medical field, judicial system etc. few words are used, which are considered as slang words in other aspects. These situations should be handled as special cases.

REFERENCES

[1] A.Gelbukh, "Computational Processing of Natural Language: Tasks, Problems and Solutions," Congreso Internacional de Computacion en Mexico D.F., Nov 15-17, 2000.

[2] George A.Miller, "WordNet: A Lexical Database," Comm. ACM, Vol.38, No.11, pp. 39-41, 1993.

[3] E. Brill, J. Lin, M. Banko, S. Dumais, A. Ng, "Dataintensive Question Answering," In: Proc. of the Tenth Text Retrieval Conference TREC-2001.

[4] Peter F. Brown, John Cocke, A. Stephen, Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to Language translation. Computational Linguistics," 16(2):79—85, 1990.

[5] A. Gelbukh, I.A. Bolshakov,"Internet, a true friend of translator," International Journal of Translation, Vol. 15, No. 2, pp. 31–50, 2003.

[6] Gelbukh, A., I.A. Bolshakov: Internet, a true friend of translator: the Google wildcard operator. International Journal of Translation, Vol. 18, No. 1–2, 2006, pp. 41–48.

[7] J. Heflin and J. Hendler, "A Portrait of the Semantic Web in Action," IEEE Intelligent Systems, vol. 16(2), pp. 54-59,2001.

[8] Y. Wilks, D. Fass, C. Guo, J. McDonal, T. Plate and B. Slator," Providing Muchine Tractablle Dictionan Tools. In Semantics and the lexicon," (Pustejowsky J. Ed.) 341-401, 1993.

[9] S. Banerjee, T. Pedersen, "An adapted Lesk algorithm for word sense disambiguation using WordNet," In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February, 2002

[10] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine conefrom an ice cream cone," Roc. 1986 SIGDOC Conference, ACM 24-26, New York of Sheffield, UK.

[11] R. Bunescu, "Associative Anaphora Resolution: A Web-Based Approach," In: Proc. of the EACL-2003 Workshop on the Computational Treatment of Anaphora, Budapest, Hungary, April.

[12] I.A. Bolshakov, A. Gelbukh,"Synonymous Paraphrasing Using WordNet and Internet," Lecture Notes in Computer Science N 3136, Springer, pp. 312–323, 2004.

[13] I.A. Bolshakov, S.N. Galicia-Haro, A. Gelbukh,"Detection and Correction of Malapropisms in Spanish by means of Internet Search," Lecture Notes in Artificial Intelligence N 3658, Springer, pp. 115–122, 2005.

[14] I.A. Bolshakov, E.I. Bolshakova, A.P. Kotlyarov, A. Gelbukh,"Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power," CICLing 2008, Lecture Notes in Computer Science N 4919, Springer, 64-72 April, 2008.

[15] A. M. Deroualt and B. Merialdo , " Natural Language modeling for phoneme-to-text transposition," IEEE transactions on Pattern Analysis and Machine Intelligence,1986.

[16] A. Ratnaparkhi, "A maximum entropy Part-of-speech tagger," Proceedings of the Empirical Methods in NLP conference, University of Pennsylvania, 1996.

[17] M. Nameh, S.M. Fakhrahmad, M. Zolghadri Jahromi,"A New Approach to Word Sense Disambiguation Based on Context Similarity," Proceedings of the World Congress on Engineering 2011 Vol I).

[18] SUO Hong-Guang, LIU Yu-Shu, CAO Shu, CAO Shu-Ying,"A Keyword Selection Method Based on Lexical Chains," Journal of Chinese Information Processing, 2006, 20(6):25-30.

[19] Jason C. Hung, Ching-Sheng Wang, Che-Yu Yang, Mao-Shuen Chiu, George YEE, "Applying Word Sense Disambiguation to Question Answering System for E-Learning," Proceedings of the 19th Inter. Conf.on Advance Information Networking and Applications (AINA'05).

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "WordNet An on-line lexical database," International Journal of Lexicography, 3(4):235-244, 1990.

ISBN: 978-81-923777-9-7