Evaluation of Authorship Attribution Software on a Chat Bot Corpus

Nawaf Ali
Computer Engineering and Computer Science
J. B. Speed School of Engineering
University of Louisville
Louisville, KY. USA
ntali001@louisville.edu

Musa Hindi
Computer Engineering and Computer Science
J. B. Speed School of Engineering
University of Louisville
Louisville, KY. USA
mmhind01@louisville.edu

Roman V. Yampolskiy
Computer Engineering and Computer Science
J. B. Speed School of Engineering
University of Louisville
Louisville, KY. USA
roman.yampolskiy@louisville.edu

Abstract— Authorship recognition is a technique used to identify the author of an unclaimed document, or in case when more than one author claims a document. Authorship recognition has great potential for applications in Computer forensics. The intended goal of this research is to identify a Chat bot by analyzing conversation log files. This is a novel area of investigation, as artificially intelligent authors have not been profiled based on their linguistic behavior. The collected data comes from chat logs between different Chat Bots and between Chat Bots and Human users. The initial experiments utilizing collected data demonstrate the feasibility of our approach.

Keywords- Authorship attribution, Authorship recognition, Chat bot, JGAAP, Stylometry.

I. INTRODUCTION

A significant amount of research has been done in the area of authorship attribution [1, 2]. Stylometry is the study of differentiating authors by their styles; a survey of the field has been presented by Stamatatos [3]. Four main methods of authorship identification are: lexical, syntactic, semantic, and content specific. In lexical methods, the word counts and distributions in the text to grasp more knowledge about the different kinds of statistical properties in the document. Syntactic methods focus on extracting specific features from the document and trying to use them to differentiate between documents. The semantic and content-based methods focus on vocabulary and choice of specific words in the document [4].

Behavioral traits associated with each human give a way to identify the person by the biometric profile [5]. Certain characteristics pertaining to language, composition, and writing, such as particular syntactic and structural layout traits, patterns of vocabulary usage, unusual language usage, and stylistic traits remain relatively constant. Identifying and learning these characteristics is the main challenge for authorship identification [6].

Two major subfields of the authorship attribution are:

- Authorship Recognition: In which we have more than one author claiming a document. We need to decide who is the best candidate to be the correct author of the document after analyzing the document and comparing it with the author's baseline profile.
- Authorship Verification: In which we have an author claiming the ownership of a document, and we need to compare the author's profile and the document analysis to see if he or she is the real author.

A good example of the first type is the twelve Federalist papers claimed by both Alexander Hamilton and James Madison [7]. A good example of the second type is a threat letter. This type is mostly used in Forensic investigation.

When talking about humans, a major challenge is that the writing style during the professional career of the writer might evolve and develop with time, a concept known as behavioral drift [8]. A similar problem arises with Chat bots, as they learn new styles of writing, will it still be possible to linguistically profile them?

A. What is a Chat bot?

A Chat bot, Chatter bot, Chatter box or Chatter robot is a computer application designed to simulate a conversation with a human user [9]. Chat bots are mainly used in applications such as online help, e-commerce, customer services, call centers, and internet gaming [10].

Chat bots are typically perceived as engaging software entities, which humans can talk to. Some Chatter bots use sophisticated Natural Language Processing Systems (NLPS), but many just scan for keywords within the input and pull a reply with the most matching keywords [11]. Chat bots are still a largely developing technology; consequently, quality of simulated conversations varies from realistic to mostly nonsense.

B. Motivations:

The online criminal community is utilizing Chat bots as a new way to steal private information, commit fraud and identity theft. The need for identifying Chat bots by their

978-1-4577-0746-9/11/$26.00 ©2011 IEEE

style is becoming essential to overcome the danger of online criminal activities.

II. Application and Data Collection

A. *Application for collecting chat logs.*

A C# application was developed to connect two Chat bots from dozens available online and to start a chatting session between them for some time. The application saves three different text files for each session: one for the whole conversation, another one for the first Chat bot part, and the last file is the other Chat bot lines. Fig 1 presents the flow chart of the application used to collect the data.

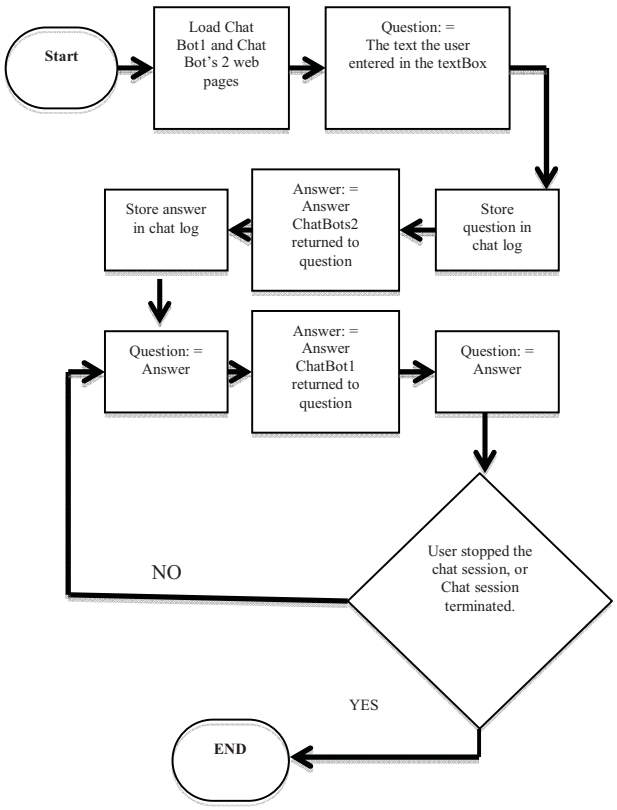

Fig. 1. The Chat bot Reloaded Application Flow Chart.

To collect additional data, chat logs on Chat bot's websites were used. Usually these logs were of chats with humans for the purposes of bot training.

III. DATA PREPARATION

The collected data had to be cleaned first from some labels like the Chat bot name, and time-date of conversation (Fig. 2). When using the chat logs taken from the Chat bot website, usually the file had both parties of the chat, so we had to remove the other person's responses.

Two applications were used on the data. The first was Stylometry [12] which was originally built for email authorship attribution. The versions available online didn't show promising results and no further tests were made. The second application was the Java Graphical Authorship Attribution Project (JGAAP) [13]. It has a practical interface with very flexible options to choose from in each of its processing stages after loading the document set. The application allows you to state the author for each file or leave it as unknown (Fig.3), and save the set as a corpus.

Jeeney: hi. How are you today? **SkyNet: I am excellent this refreshing winter day.** **Jeeney: What do you think of me?** **SkyNet: I have not come to a conclusion about what of me.** **Jeeney: Why didn't you?** **SkyNet: I thought it was the right course of action.** **Jeeney: Why did you think that?** **SkyNet: It is part of who I am.** **Jeeney: Oh, okay.**	**TalkBot: Hi, I am TalkBot.** **Hal: I have never heard of that name. What nationality is TalkBot?** **TalkBot: Not that I recall. I often hear many things from people who drop by to chat but I sure don't remember it all.** **Hal: Not many people express themselves that way.** **TalkBot: Perhaps, but a significant number nevertheless.** **Hal: Sounds okay.**

Fig. 2. Sample conversation between Chat bots.

Fig. 4 illustrates five main available options:

- Normalize white space
- Null canonicize (No preprocessing)
- Strip alpha numeric
- Strip punctuation
- Unify case

One can add one or more preprocessing to all files or to any selected files. The next stage was the event set. It is more like the feature extraction stage; in which one can select what kind of feature to extract from the text files in this stage as shown in Fig 5.

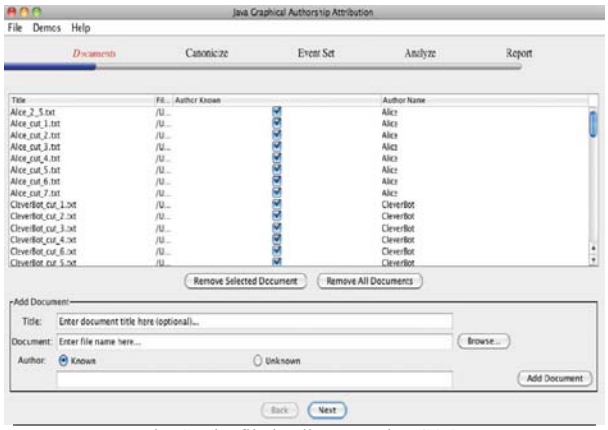

Fig. 3. The file loading stage in JGAAP.

After selecting the feature extraction step, we step forward to the analyze stage or the classifier selection. One can choose from different classifiers to test on the data (Fig. 6). The last stage is the report and the result of classification. One will have a list of all the files with the

unknown author with the attributed author next to it (Fig. 7).

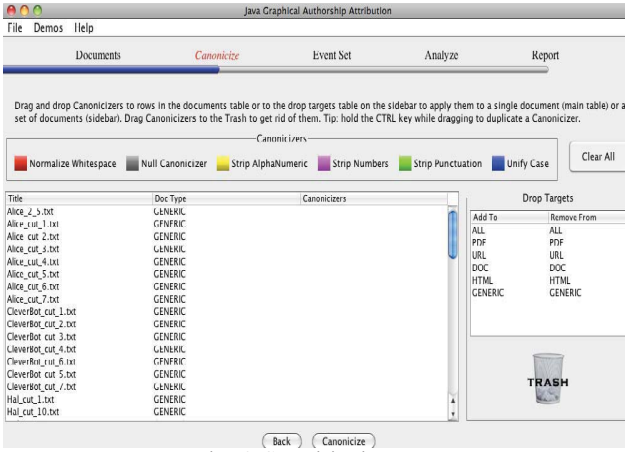

Fig. 4. Canonicization stage.

Fig. 5. Event Set or the Feature Extraction stage.

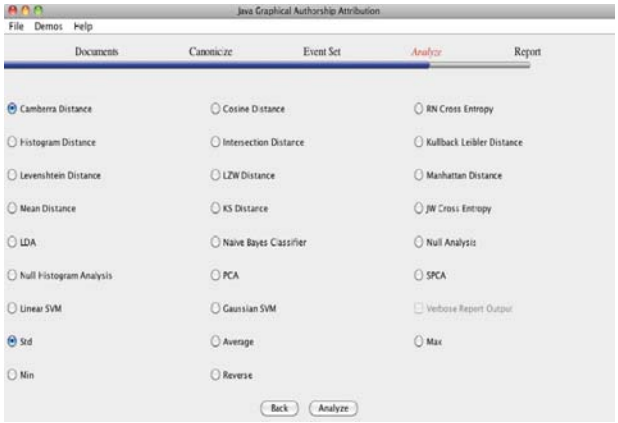

Fig. 6. Analyze stage or Classifier selection.

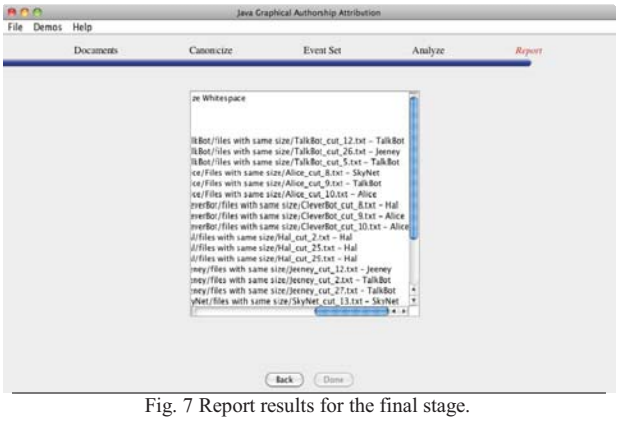

Fig. 7 Report results for the final stage.

IV. CHAT BOTS USED.

Eleven Chat bots were used in the initial experiments: Alice [14], CleverBot [15], Hal [16], Jeeney [17], SkyNet [18], TalkBot [19], Alan [20], MyBot [21], Jabberwock [22], Jabberwacky [23], and Suzette [24].

As mentioned before, data was collected by different means; some from Chat bot sites themselves, other by letting two Chat bots talk to each other. Fig 8 shows some of the Chat bots used.

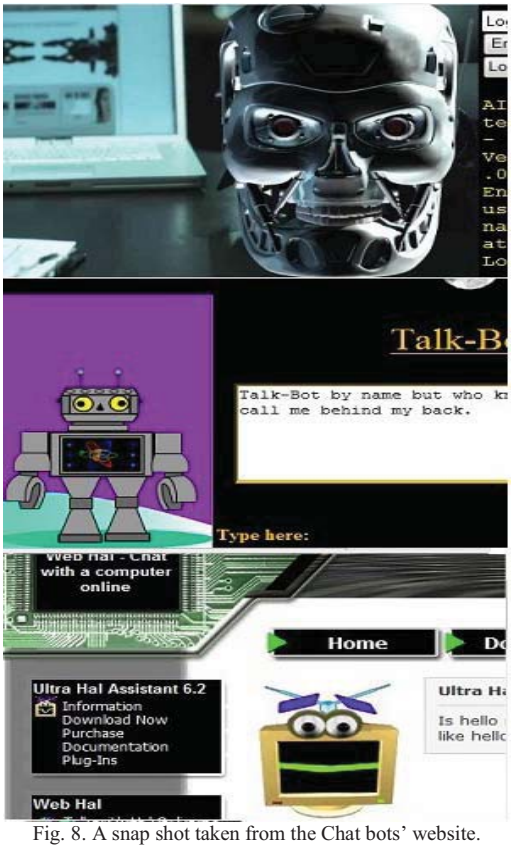

Fig. 8. A snap shot taken from the Chat bots' website.

TABLE I. AVERAGE ACCURACY FOR EACH CLASSIFIER USING DIFFERENT FEATURES

Classifier Used	Average Accuracy using different Features per Classifier
JW Cross Entropy	72.05%
KS Distance	65.34%
Camberra Distance	65.18%
Cosine Distance	65.11%
Histogram Distance	65.11%
Manhattan Distance	64.72%
Kullback Leibler Distance	59.80%
Levenshtein Distance	45.04%
Intersection Distance	42.54%
LDA	42.08%
RN Cross Entropy	39.27%
Naïve Bayes Classifier	16.32%
LZW Distance	12.49%
Mean Distance	0.00%

V. EXPERIMENTS

For each type of preprocessing, different event sets and classifiers were tested (Fig. 9). A total of 306 different tests were conducted on the data set.

Fig 10 and 11 shows the results from the experiments conducted on the available data. It also shows how the data interacts with different selections of features when applying different classifiers. Table 1 shows the average accuracy for each classifier when using different features.

The Juola & Wyner Cross Entropy classifier achieved the maximum accuracy [13]. The only drawback for this algorithm was speed; it was very slow to do the classification with JW Cross Entropy. Table 2 shows the average accuracy achieved for each Feature over all classifiers used. Maximum accuracy was achieved by the Vowel 2-3 letters words feature (60.30%), which finds all the words starting with vowels with a length of 2-3 letters. The minimum accuracy was obtained by using the Dis Legomena feature (20.96%). This feature calculates words appearing only twice in the document.

Table 3 explains the features used and their usage description.

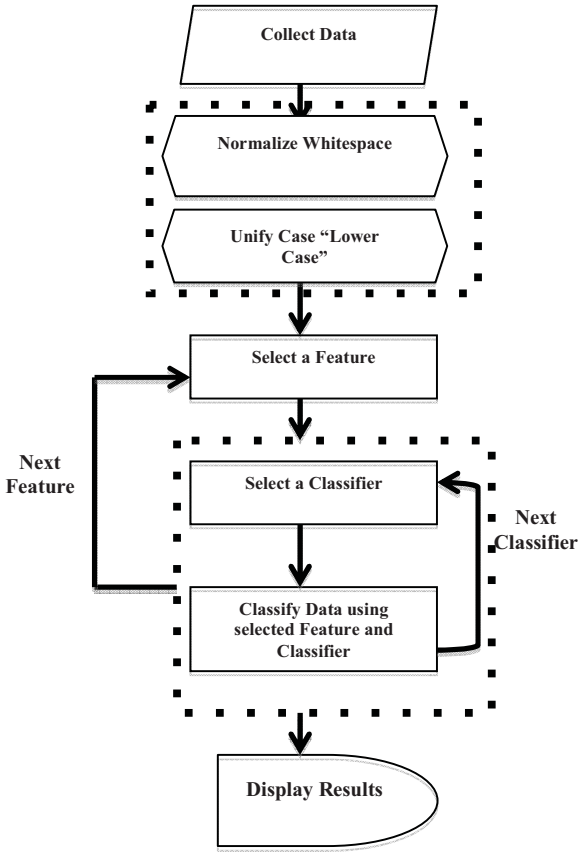

Fig.9. Flow Chart of the process followed during the Experiments.

TABLE 2. AVERAGE ACCURACY FOR EACH FEATURE OVER ALL CLASSIFIERS.

Feature Used	Average per Feature	Feature Used	Average per Feature
Vowels 2-3 letters words	60.30%	MW function Words	52.46%
Vowels 2-4 letters words	59.48%	Word Bigrams	51.87%
2-4 Letters	59.13%	Vowels 3-4 letters word	49.77%
Vowel initial words	58.31%	Word Length	39.58%
2-3 Letters	58.20%	Word Trigrams	38.64%
Character Bigrams	57.26%	Syllables per word	28.57%
Characters	56.56%	Hapax-Dis Legomena	24.36%
Character Trigrams	55.74%	Word Tetra Grams	24.36%
Words	55.39%	Hapax Legomena	22.83%
Character Tetra Grams	55.27%	Dis Legomena	20.96%
3-4 Letters	53.51%		

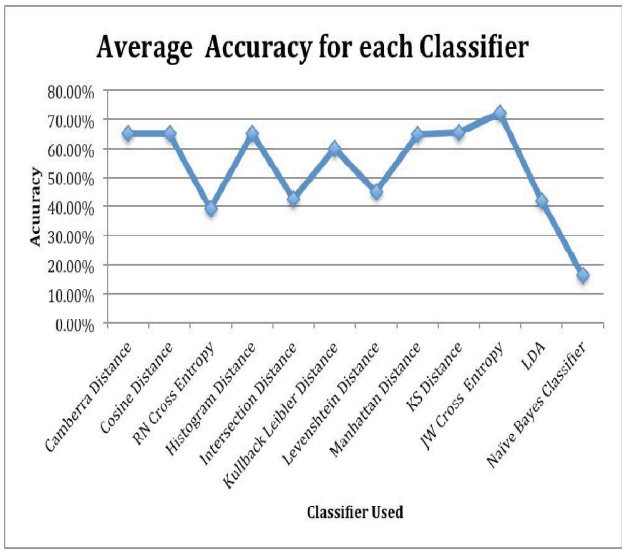

Fig 10. Average accuracy for each classifier

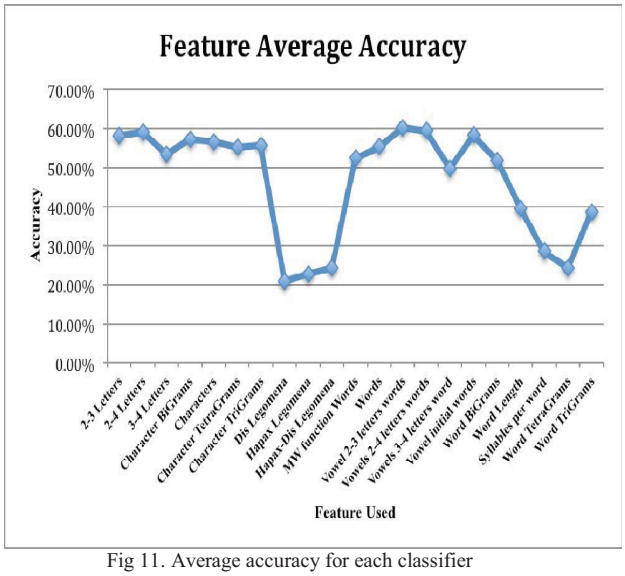

Fig 11. Average accuracy for each classifier

VI. Conclusions and Future Work

Our experiments in profiling Chat bots by their linguistic behavior using authorship identification techniques did show the feasibility of such approach.

JGAAP [13] has given us the ability to test each feature's performance on our data. Trials showed that

TABLE 3. FEATURES' DESCRIPTION

Feature Used	Feature usage description
2-3 Letters	Words with 2 or 3 letters length.
2-4 Letters	Words with 2,3, or 4 letters length.
3-4 Letters	Words with 3 or 4 letters length.
Character Bigrams	Character pairs in sequence.
Characters	Unicode Characters frequencies.
Character Tetra Grams	Groups of four successive letters.
Character Trigrams	Groups of three successive letters.
Dis Legomena	Words appearing only twice in the document.
Hapax Legomena	Words appearing only once in the document.
Hapax-Dis Legomena	Words appearing once or twice in the document.
MW function Words	Function words from Mosteller-Wallace.
Words	Words frequencies (white space as separator).
Vowels 2-3 letters words	Words starting with a vowel with length of 2 or 3 letters.
Vowels 2-4 letters words	Words starting with a vowel with length of 2, 3, or 4 letters.
Vowels 3-4 letters word	Words starting with a vowel with length of 3 or 4 letters.
Vowel initial words	Words starting with a Vowel (A, E, I, O, U).
Word Bigrams	Word pairs in sequence.
Word Length	The length of words in each document.
Syllables per word	Number of vowel cluster per word.
Word Tetra Grams	Groups of four successive words.
Word Trigrams	Groups of three successive words.

'Vowels 2-3 Letter words' feature resulted in the highest average accuracy of (60.30%). Juola and Wyner Cross Entropy resulted in the highest average accuracy of (72.05%) but was slow. The only drawback was that we couldn't try more than one feature at a time. The next step is to find out the best combination of features that can perform well on the Chat bot data set, and test the 'Stacking/Bagging' technique of using more than one classifier on the data set to obtain higher classification accuracy.

References

[1] E. Stamatatos, "Text Sampling and Re-sampling for Imbalanced Authorship Identification Cases," in *17th European Conference on Artificial Intelligence(ECAI), August 29 - September 1, 2006*, Riva del Garda, Italy, pp. 813-814.

[2] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group Web forum messages," *IEEE Intelligent Systems,* vol. 20(5), pp. 67-75, 2005.

[3] E. Stamatstos, "A survey of modern authorship attribution methods," *Journal of the American Society for Information Science and Technology,* vol. 3, pp. 538-556, 2008.

[4] R. H. R. Tan and F. S. Tsai, "Authorship Identification for Online Text," in *International Conference on Cyberworlds (CW)*, Singapore, Singapore, 2010, pp. 155-162.

[5] R. V. Yampolskiy and V. Govindaraju, "Behavioral Biometrics: a Survey and Classification," *International Journal of Biometrics (IJBM).,* vol. 1(1), pp. 81-113, 2008.

[6] O. Angela, "An Instant Messaging Intrusion Detection System Framework: Using character frequency analysis for authorship identification and validation," in *40th Annual IEEE International Carnahan Conferences Security Technology*, Lexington, KY, 2006, pp. 160-172.

[7] D. I. Holmes and R. S. Forsyth, "The Federalist Revisited: New Directions in Authorship Attribution," *Literary and Linguistic Computing,* vol. 10(2), pp. 111-127, January 1, 1995 1995.

[8] M. B. Malyutov, "Authorship attribution of texts: a review," *Electronic Notes in Discrete Mathematics,* vol. 21, pp. 353-357, 2005.

[9] Chatbot. (2011, June, 2011). *Chatbot - Artifficial person with interactive textual conversation skills*. Available: www.chatbot.org/chatbot/

[10] Webppedia. (June, 20, 2011). *What is chat bot? A Word Definition from the Webpedia Computer Dictionary*. Available: www.webopedia.com/TERM/C/chat_bot.html

[11] Wikipedia. (2011, June, 22). *Chatterbot-Wikipedia, the free encyclopedia*. Available: www.en.wikipedia.org/wiki/Chatterbot

[12] Pace_University. (2011, June, 4th). *Stylometry*. Available: http://utopia.csis.pace.edu/cs615/2006-2007/team2/

[13] P. Juola. (2011, July, 4th). *Java Authorship Attribution Application*. Available: http://evllabs.com/jgaap/w/index.php/Main_Page

[14] ALICE. (2011, June, 12). *ALICE* Available: http://alicebot.blogspot.com/

[15] CleverBot. (2011, July, 5th). *CleverBot* Available: http://cleverbot.com/

[16] HAL. (2011, June, 16th). *AI Research*. Available: http://www.a-i.com/show_tree.asp?id=97&level=2&root=115

[17] Jeeney. (2011, March, 11). *Artificial Intelligence Online*. Available: http://www.jeeney.com/

[18] SkyNet. (2011, April, 20). *SkyNet - AI*. Available: http://home.comcast.net/~chatterbot/bots/AI/Skynet/

[19] TalkBot. (2011, April, 14th). *TalkBot- A simple talk bot*. Available: http://code.google.com/p/talkbot/

[20] Alan. (2011, June, 10). *AI Research*. Available: http://www.a-i.com/show_tree.asp?id=59&level=2&root=115

[21] MyBot. (2011, Jan,8). *Chatbot Mybot, Artificial Intelligence*. Available: http://www.chatbots.org/chatbot/mybot/

[22] Jabberwock. (2011, June, 12). *Jabberwock Chat*. Available: http://www.abenteuermedien.de/jabberwock/

[23] Jabberwacky. (2011, June, 10). *Jabberwacky-live chat bot-AI Artificial Intelligence chatbot*. Available: http://www.jabberwacky.com/

[24] Suzette. (2011, Feb, 7). *SourceForge ChatScript Project*. Available: http://chatscript.sourceforge.net/