BlogWall: Word Sense Disambiguation based topic summarization and genetic algorithms to generate better poem remixes based on user input.

Vidyarth Eluppai Srivatsan

Department of Electrical and Computer Engineering
National University of Singapore

Abstract—**The younger generation today is fast embracing popular culture and this has inadvertently endangered the preservation of intangible cultural sources like traditional poetry. We attempt making it an integral part of the consuming popular culture, thereby introducing the topic of preservation to the younger generation. However, it is difficult for most of us to actually create a poem. The BlogWall is an attempt to bridge this gap. By blending media art and poetry, we have developed a poetry mixer called to extend text messages to a new level of self expression and public communication. From a single text message, the system is capable of creating a new poem by drawing from an existing body of poetry. The technique integrates a number of ideas from different disciplines such as information retrieval and natural language understanding specifically word sense disambiguation & topic summarizing, and augments the system with basic genetic algorithms to create a model for coherent and relevant poetry generation and remixing. Thus, it allows for new forms of cultural computing.**

I. Introduction

PEOPLE long for means of public expression. Graffiti is a one form of public expression that was very popular in mid 70s. Blogs enable people to share their ideas with millions of others all over the world. Today Internet websites such as “Youtube” are popular ways for sharing videos. Creating novel ways for people to express themselves is ever so important.

Technological development also poses a certain danger to us that it will distance us from our cultures. Younger generations interacting with these new technologies are getting absorbed into the popular cultures very fast. Literary arts such as poetry are less interesting to them. [1] Many of them would not go through the literary work such as poetry just for the joy of it [2]. This is not an inherent result of technological development, but rather the use of it. If traditional literary work can be wrapped around modern technology, the younger generation will be able to understand and appreciate it while enjoying the works of art like many generations before them. There is tremendous potential for new forays in the realms of cultural computing.

Poetry is deeply rooted in traditional culture, and that is what we will re-capture with our system. Furthermore, poetry is a great way of expressing inner thoughts. However, the inherent characteristics of poetry render traditional approaches ineffective. For instance, poems generally do not have clear and well-defined communication goals. They rely rather on abstract and figurative language, encouraging the reader to form their own conclusions as to their meaning.

Essentially, most poetry generation so far has consisted of randomly choosing words and making the resulting phrases fit in a predefined language grammar. Natural language understanding that aims to mimic human-like communication between man and machine is also inadequate on its own when it comes to generating poetry.

The BlogWall poetry mixer combines natural language processing methods like word sense disambiguation, topic summarizing and genetic algorithms to generate poetry. It is an attempt to recreate social communication among the youth by drawing from the phenomena of “mixing” or “mash-up” and applying these ideas to poetry. The novel interface to poetry as well as wide usage of text messages among younger generation would make this application very appealing. With this effort we hope to create new form of text art as well as attract the younger generation to literary works such as poetry.

II. Related Work

Today we continue to build a variety of applications that provide entertainment to the younger generation but only a limited number of them actually combine art and culture to enrich their experience.

A. *Art, Public Displays and Messaging*

One of the pioneering works in cultural computing was ZENetic computer [3]. It is an interface that evokes self-awakening through important aspects of Zen Buddhist culture. It tries to offer users a chance to engage and understand Buddhist principles of ‘recreation’ of the self. With stories portrayed in ink, haiku, and kimono, the ZENetic conveys the rich allegorical interactive characteristic of Eastern philosophy.

Researchers around the world have being experimenting with different combinations of art, public displays and mobile messaging. The mobile phone has already been used as a medium of self expression [4]. Ballagas et al. [5] discuss enabling interactions with large public displays using mobile phone. They have used the embedded camera on mobile phones as an enabling technology. Ballagas et al.’s “Point & Shoot” technique allows users to select objects using visual codes to set up an absolute coordinate system on the display surface instead of tagging individual objects on the screen. Joe Blogg [6] is a public display where users can contribute content by sending messages and images to it using their mobile phones. TexTales [7] is

The initial version of the Blogwall was funded by Singapore Science Centre. This project is installed in the iSpace at Singapore Science Centre.

a large-scale photographic installation to which people can send text message captions. It can create technologically supported public discourse spheres in which they can both represent personal views and practice new ways of forming collective opinions. Mobile phone can also act as a controller of a public display, for example in the Blinkenlights [8] project the upper eight floors of the building were transformed in to a huge display by arranging 144 lamps behind the building's front windows. By using mobile phones, users could play a game of “Pong”. One of the pioneering works in cultural computing was “Hitch Haiku”, which automatically generated Haiku from a database of books [11]. The BlogWall consists of many of the features found on those systems but it concentrates on promoting artistic and social communication through poetry.

B. *Poetry Generation and Remixing*

“i.plot" [10] is a system that discovers hidden connections between unrelated words by tracing possible paths through a database, traversing many two-word connections built from content based on publicly available resources. Our model extends the concept of “i.plot" further by making connections between the input text and the poems, as well as among the poem lines themselves.

So far, attempts at generating language prose have essentially been in a similar vein as PROSE or RACTER [11]; two examples that exist in publication. These are in turn similar to ELIZA [12] and FRED [13], in their approach, which essentially consists of creating prose at random but suited to a grammar template. However, these methods are inadequate since they neither account for the inherent abstract nature of poetry nor for the rhythm and timing of poems. In view of these limitations, Manurung et al. [14] proposed the “Stochastic Hillclimbing Model” which attempts to address these difficulties.

The poetry generation model in the BlogWall consists of several stages. The system uses three different criteria to shortlist discrete sets of poem lines. The focus of our proposed poetry remixing model is the ability to form meaningful connections between an existing body of poetry and the input text from the user. By drawing from the existing body of work, the BlogWall’s poetry remixing algorithm will then attempt to create an original poem which is meaningful and entertaining to the user.

III. Preliminaries

The BlogWall maintains 2 individual components that are integrated into the main system to give good quality poem mixing. We shall discuss those in the following sections.

In the discussion that follows, we use the following terminology.

A. *Word Senses/ Concepts & Word Sense Disambiguation (WSD)*

Concepts are real world objects, notions or ideas that are represented in text or speech by words. For example, the concept of a stone would be represented by the word stone. In addition, it may well be represented by the word rock or pebble. Hence, the same concept may be represented by different words. Also, the concept need not be a solid object. It could be an abstract thing, like art, or an action, like walking. Each such concept has a number of words that represent it. Not only that, but a single word may represent a number of concepts. For example, the word bank could mean the financial institution concept or the river bank concept. The different meanings of a word are known as word senses. A word could, therefore, correspond to a number of concepts, while a word sense corresponds only to a single concept. Due to this equivalence of word senses and concepts, in this thesis we use the terms concepts and word sense interchangeably. Word Sense Disambiguation is the process of assigning a meaning to a word based on the context in which it occurs. [15]

B. *Semantic Relatedness & Similarity between 2 words*

Semantic relatedness implies how closely two words/ concepts are related through relationships like antonyms, synonyms etc. Semantic similarity on the other hand represents how alike 2 words/ concepts are. For WSD we use relatedness.

C. *WordNet and its related terminology*

WordNet is a lexical database [16]. WordNet can be visualized as a large graph or semantic network, where each node of the network represents a real world concept. For example, the concept could be an object like a house, or an entity like a teacher, or an abstract concept like art, and so on.

Every node consists of a set of words, each representing the real world concept associated with that node. Thus, each node is essentially a set of synonyms that represent the same concept. WordNet also has various relating links between the synsets. For example, relationships of the form “a vehicle is a kind of conveyance” or “a spoke is a part of a wheel” are defined. Other relationships include is opposite of, is a member of, causes, pertains to, etc. On relation that we use a lot in this paper is “is a” hierarchy relation. For example a man is a person. Thus person is an hypernym (the is a hierarchy) of man.

D. *Term Frequency*

The term frequency is a measure of how often a particular term occurs in a poem lines. For instance, if the word ‘beautiful’ appeared twice in a poem line consisting of 10 words, the term frequency of the word ‘beautiful’ would be 0.2. Mathematically, this can be described as follows:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (1)$$

where $n_{i,j}$ is the number of occurrences of the term being considered in document d_j, and the denominator is the number of occurrences of all terms in document d_j .

E. Super Sense & Specific Topic

The general root sense of a word (e.g.) man belongs to the super sense person. In WordNet terms super sense is the upper level hypernym of a word (not the uppermost) Specific Topic is a specific case of which the super sense is a hypernym and occurs the most frequently in the content for that particular super sense and order by the furthest distance away from the super sense hypernym. This cannot be a pronoun. Man is a possible specific topic for person. Please note the terms **topic** and **super sense** are used interchangeably in this thesis.

F. Salience Measure of a particular super sense (or super sense less specific topic) in a document

The **salience measure** is a combination of the following factors (in decreasing order of importance)

1. As Subject Term Frequency
   Term frequency of the index as a subject. (Subjects are found by a very simple parse tree algorithm)
2. Syntactically related Subjects frequency
   Summation of the term frequencies of the Subjects to which the index is syntactically related to over the total number of nouns. (Syntactic relation is measured by whether it occurs in conjunction with the subject)
3. Syntactically related Noun frequency
   Summation of the term frequencies of the nouns to which the index is syntactically related to over the total number of nouns. (Syntactic relation is measured by whether it occurs in conjunction with the noun)
4. Term Frequency
   Simple term frequency of the index

Also the importance for various types of super senses (words) is different while summarizing. The decreasing order of importance of word types is as follows,

1. Nouns (includes pronouns & prepositions)
2. Verbs, Adverbs and Adjectives
3. Auxiliary Verbs, Connectors (excluding just coordinating & super sense less ones) and Prepositions

The above is the order in which results from the first set are ordered into.

IV. Measuring Semantic Relatedness & Word Sense Disambiguation

Before going further, a quick note on why we need this part. Word Sense Disambiguation is needed to understand the exact sense of a word and thereby it helps in topic summarization, finding similarity between words both semantic and syntactic. This is essential when we want to remix poems as it is these measures that enhance information retrieval both qualitatively and quantitatively & summarization of it. Simple term frequencies are very limited because it uses very little intra document statistics or simply the context.

Measuring the semantic relatedness of concepts is an intriguing problem in Natural Language Processing. Various approaches that attempt to approximate human judgment of relatedness, have been tried by researchers. In this section we compare the existing standards for measuring semantic relatedness.

Referring to Ted Pederson's article [15] where he compares the methods

- Jiang & Conrath Measure
- Resnik Measure
- Lesk Measure
- Hirst & St Onge Measure
- Lin Measure
- Leacock & Chodorow Measure

we infer that Jiang & Conrath Measure and Lesk Measure perform the best. Also Pedersen notes that neither path length measure nor Wu & Palmer Measure comes anywhere close to any of the aforementioned measures. But we note that the Vector measure suggested by Siddharth [16] combines elements of Information Content methods used by Jiang & Conrath & Glosses method used by Lesk measures. So we sought out to compare these 3 methods for our poem database and input message database (over 500 words).

Table 1: Performance of the best relatedness measures for our sample set

Method	*% unidentified*	*% identified Correctly (out of those identified)*
Vector	**50.00%**	70.00%
Lesk	**50.00%**	55.00%
Jcn	**50.00%**	45.00%

We observed that over 50% of the words were unrecognized. This was found to be the fault of not the measures but rather of WordNet, which didn't have the common words like "the, a, an" etc stored. This was a huge problem because the word sense disambiguation layer forms the basis for the topic summarizing layer. With such a low accuracy our whole algorithm will suffer. Hence the need to modify these algorithms arose. We then supplemented some dictionary based features to WordNet to handle the following

1) Missing pronoun handling
2) Missing conjunction handling
3) Missing preposition handling
4) Missing auxiliary verb handling
5) Missing punctuation especially quotes handling

We built our solution on top of the vector measure as it was the best of those measures currently available. Here is the result that we obtained for our algorithm for the sample set used as above for the other measures

Table 2: Performance of our modified vector measure for our sample set

Method	*% unidentified*	*% identified Correctly (out of those identified)*
Modified Approach	**5.00%**	**85.00%**

The number of words that went unidentified reduced to almost 5% from 50%, a tenfold reduction. Another point to note is that even the accuracy of the algorithm for the identified words has increased by 15%. This is essentially due to the pronoun handling and the punctuation handling which cuts down on mistakes and enhances accuracy.

An accuracy of 85% to understand words & 95% to identify words keeps us in good stead for the other upcoming components – topic summarizing & genetic coherence generator.

V. Topic Summarization

We need to use topic summarization because that will help us select poem lines that are more relevant and do that faster.

Here is a brief overview of the process

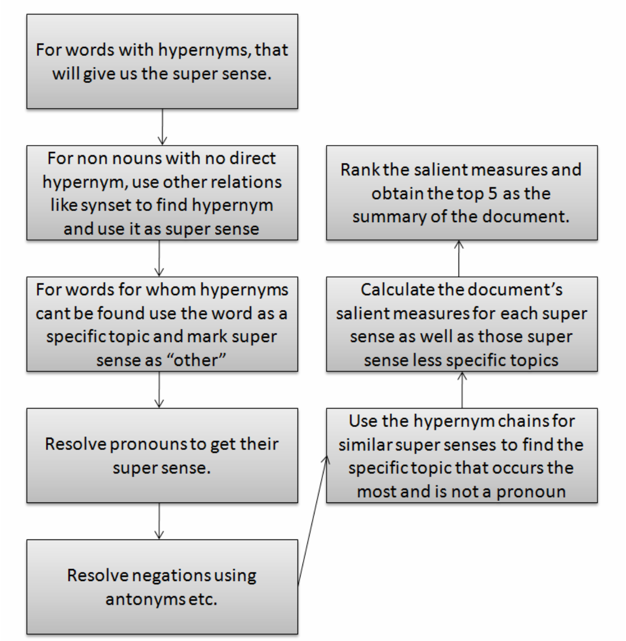
For words with hypernyms, that will give us the super sense.
For non nouns with no direct hypernym, use other relations like synset to find hypernym and use it as super sense
For words for whom hypernyms cant be found use the word as a specific topic and mark super sense as “other”
Resolve pronouns to get their super sense.
Resolve negations using antonyms etc.
Use the hypernym chains for similar super senses to find the specific topic that occurs the most and is not a pronoun
Calculate the document’s salient measures for each super sense as well as those super sense less specific topics
Rank the salient measures and obtain the top 5 as the summary of the document.

Figure 1: Topic Summarization Process

Hypernyms are defined as the parent of an “is a” hierarchical relationship. For example person is an hypernym of a man as a man is a person.

We did some preliminary testing on 20 set of text files each with at least a few hundred words. And here are the results

Table 3: Accuracy of Topic Summarizing

nth Topic	*Accuracy %*
1	**90.00%**
2	80.00%
3	65.00%
4	50.00%
5	35.00%

The above results are expected; as the salience measure goes down then the accuracy of that being a topic is reduced. This accuracy in topic summarizing is good enough for us to apply genetic algorithms based on it.

VI. Genetic Algorithms To Enhance Coherence Of Remixed Poems

We need genetic algorithms to enhance the quality of the poems generated. A brief overview of this process is shown below.

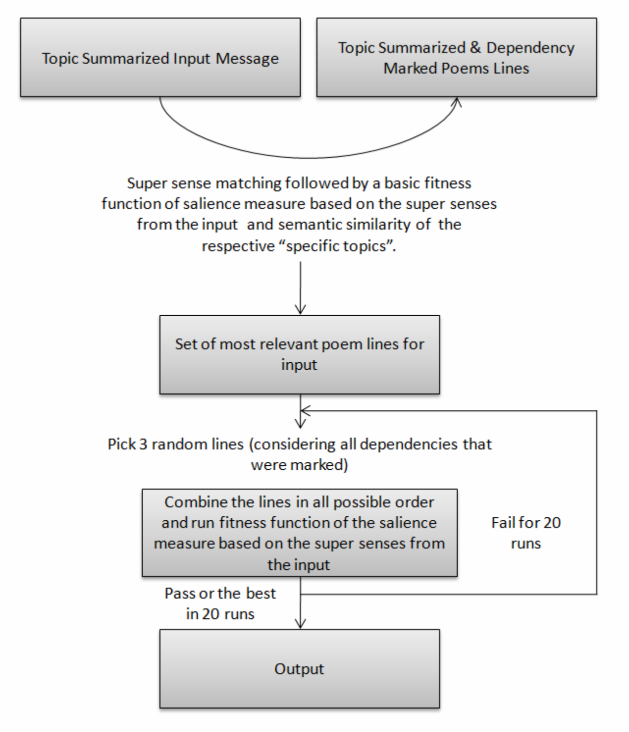
Topic Summarized Input Message
Topic Summarized & Dependency Marked Poems Lines
Super sense matching followed by a basic fitness function of salience measure based on the super senses from the input and semantic similarity of the respective “specific topics”.
Set of most relevant poem lines for input
Pick 3 random lines (considering all dependencies that were marked)
Combine the lines in all possible order and run fitness function of the salience measure based on the super senses from the input
Fail for 20 runs
Pass or the best in 20 runs
Output

Figure 2: Genetic Algorithm that utilizes WSD, Topic Summarization to create remixes

Dependency marking for the poem includes grammatical dependencies only like ‘either’ statements must be followed by ‘or’ statements etc.

Also semantic similarity is measured here (not semantic relatedness) when selecting poem lines. It is measured by a combination of the following factors

1. Part of the same synset or other similar relations.
2. Part of antonym relation but there is a negation qualifier in front of one of the *specific topic*

The above 2 factors are binary in the sense that they give a value of 0 (not found) and 1(found). The next measurement is more dynamic. It calculates the similarity of the hypernym chains till the super sense.

For example take the words ‘man’ and ‘son’.
Hypernym chain for man: Man” is a” male ”is a” person
Hypernym chain for son: Son ”is a” male ”is a” person

Hypernym chains of each of those match exactly (apart from the words). So this receives a full score for hypernym linkage factor. If let us say of 4 possible hypernym levels 2 levels match it will give us half the score for this section. We use a combination of the above 3 factors as a fitness function to understand semantic similarity of super senses in the input and the poem line. Notice that this is much deeper than just fetching the synonyms.

VII. Poetry Generation And Remixing

The poetry generation process in the poetry mixer consists of several stages. The system uses the three different components described above (Section IV, V & VI) to criteria to shortlist discrete sets of poem lines. The schematic in Figure 2 illustrates this run time process while Figure 4 illustrates how the various components fit in.

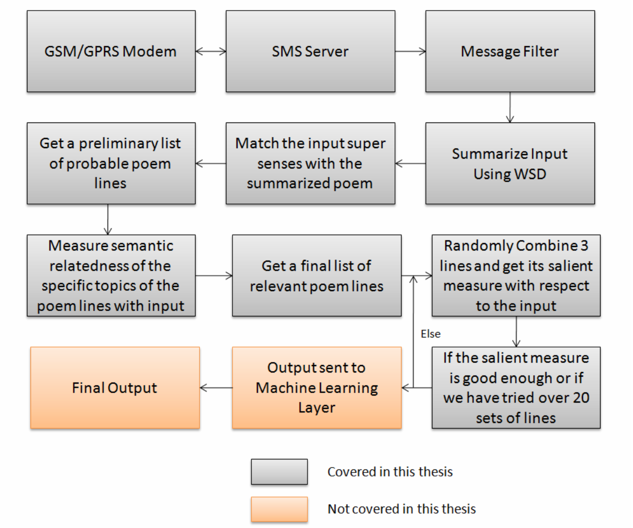
GSM/GPRS Modem
SMS Server
Message Filter
Get a preliminary list of probable poem lines
Match the input super senses with the summarized poem
Summarize Input Using WSD
Measure semantic relatedness of the specific topics of the poem lines with input
Get a final list of relevant poem lines
Randomly Combine 3 lines and get its salient measure with respect to the input
Else
Final Output
Output sent to Machine Learning Layer
If the salient measure is good enough or if we have tried over 20 sets of lines
Covered in this thesis
Not covered in this thesis

Figure 3: Poetry Generation Overview

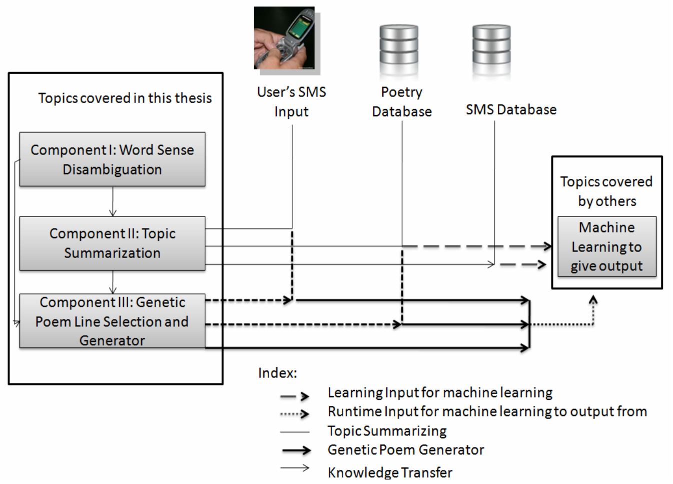
Topics covered in this thesis
User's SMS Input
Poetry Database
SMS Database
Component I: Word Sense Disambiguation
Component II: Topic Summarization
Component III: Genetic Poem Line Selection and Generator
Topics covered by others
Machine Learning to give output
Index:
Learning Input for machine learning
Runtime Input for machine learning to output from
Topic Summarizing
Genetic Poem Generator
Knowledge Transfer

Figure 4: Component Usage

Let us take an example from the new system. Let us assume the input to be “Be a strong man”. This is the eventual output from the preliminary prototype system (excluding the AI part of my fellow researcher):-

“A strong man puts God first
Be strong and wade through the impending doom
And - which is more - you'll be a Man, my son!”

Compare this to a system, which just uses term frequency and synonym matching for which if you give the input as “We are not machines” you get the output as,

“And all kinds of benevolent machines,
We melted, merged, meandered
Whereas letters are fixed.”

Clearly there seems to be a bit more coherence & meaning in the poem generated by the new poem remixer. A quick run through of how this poem was made follows. The input line’s super senses are ‘person’, ‘attribute’ and ‘other’ with specific topics ‘man’, ‘strong’ and auxiliary verb ‘be’. The lines selected had multiple overlaps for person, attribute and ‘be’, which coupled with semantic relations between ‘man’, ‘God’, ‘son’ & ‘you’ and term frequency of each of the words (including subject frequency & subject & noun syntactic relation frequency.

VIII. Results And Discussion

Preliminary results are very encouraging, and appear to be a definite qualitative improvement over the previous algorithm. People find the system entertaining, and are curious to experiment. The output of the system generally appealed to the user. However, certain key limitations of the system were also identified as outlined in the next section.

The main novelties in this system that makes it better are as follows

1) **Better Information Summarization**
this is a direct result of the Topic Summarization method. Understanding of the context gives rise to a better summary of the poem.
2) **Better information Retrieval**
Quality wise: By understanding intra document links and understanding the sense of the words the quality of poem lines selected is improved.
Quantity wise: Again by understanding the context the algorithm is able to retrieve more relevant lines than simple term frequency.
3) **Better Remixing Quality (Especially Coherence)**
By using inter line links to reinforce the input message’s theme the remixed poem is of better standards.
4) **Limited information Loss**
No stone is left unturned unlike the previous solution where only the top 3 synonyms were used to find the relevant lines.
5) **User can use keywords to remix poetry instead of a full sentence** (unlike the previous system)

Apart from the above, the work done by my fellow FYP student Sebastien incorporates Machine Learning and improves the system over time.

All these system novelties are a direct consequence of the novel components used,

1) **Highly efficient Word Sense Disambiguation** (performs better than several industry standard measures) due to the unified usage of WordNet and dictionary methods
2) **Innovative & efficient approach to topic modeling** which is quite different from the normal probabilistic models that is prevalent nowadays. Some methods of word sense disambiguation in fact use topic modeling in their process. But a highly efficient word sense disambiguation tool has helped us build a good topic modeling system.
3) **Innovative poem line selection algorithms that use genetic programs.**

IX. Observations and Limitations

We need for larger databases to make things more interesting. More importantly we have identified a certain key issue to address.

Slow WSD processing. When the document contains a lot of line the WSD part takes a lot of time to process. Although Blogwall was lucky to have small documents (sms messages) to process on run time this might affect other project when incorporated. This is because of the highly complicated measurement of sense relatedness along with the addition of the dictionary features.

X. Conclusion

Every one of us has some level of artistic or poetic ability. However, not everyone is able to create a poem. Especially, the younger generation may not have the necessary background or the knowledge to do so. The Blogwall is a technique to bridge this gap. Text messages provide the ideal basics because the technology is familiar. The interface, using a mobile phone is also within easy reach. The Blogwall allows each individual to create their own custom and unique work of personal art, by using existing work. Thus, a very personal and unique poetry mixer can be created depending on the input text.

We are currently planning to extend the system by incorporating richer user interfaces using colors, pictures, sounds and music. Furthermore, to encourage user participation, additional ways of interaction could be provided to the user. The poetry mixer could, for example, be ported to the web. This would make it a very social application where people could export and share the works they have created with others.

Another interesting possibility is to provide the ability to the users to pick genre types and authors for the poem. Examples could include a “Shakespearean poetry generator” or “Limerick generator”. This could increase the appeal of the application, and also potentially increase the effectiveness of coming up with a poem that is relevant and entertaining.

Another possibility is to explore the option of adding, editing and removing words in a line to better suit the input message something like what Manurung suggests [14]. It could also have a layer that checks for rhyme and rhythm aspects of the poem in the genetic algorithm part. Now that the Word Sense Disambiguation Layer is up with the other tools this should not be too far from actualization.

Acknowledgment

I gratefully acknowledge the valuable guidance from Dr. Adrian David Cheok from National University of Singapore and Ted Pedersen from University of Minnesota in Duluth. I also acknowledge the many useful suggestions and initial contributions of Hooman A Samani, Sebastien, Newton, Nimesha and Ken from the MXR Lab Singapore.

References

[1] Merrick, B., Poetry and pleasure, Children's Literature in Education, Vol 10, Dec 1979, 203-205

[2] Benton, M., Poetry for children: a neglected art, Children's Literature in Education, Vol 9, Sep 1978, 111-126

[3] Tosa, N., Matsuoka, S., Miyazaki, K., Interactive storytelling system using behavior-based non-verbal information: ZENetic computer. In Proceedings of the Eleventh ACM international Conference on Multimedia (Berkeley, CA, USA, November 02 - 08, 2003). MULTIMEDIA '03. ACM Press, New York, NY, 466-467, 2003.

[4] Cheverst, K., Dix, A., Fitton, D., Kray, Ch., Rouncefield, M., Saslis-Lagoudakis, G., Sheridan, J.: Exploring Mobile Phone Interaction with Situated Displays. PERMID 2005: 43-47.

[5] Ballagas, R., Rohs, M., and Sheridan, J. G. 2005. Sweepand point and shoot: phonecam-based interactions for large public displays. In CHI '05 Extended Abstracts on Human Factors in Computing Systems (Portland, OR, USA, April 02 - 07, 2005). CHI '05. ACM Press, New York, NY, 1200-1203.

[6] Martin, K., Penn, A., and Gavin, L. 2006. Engaging with a situated display via picture messaging. In CHI '06 Extended Abstracts on Human Factors in Computing Systems (Montréal, Québec, Canada, April 22 - 27, 2006). CHI '06. ACM Press, New York, NY, 1079-1084.

[7] Ananny, M., Strohecker, C., Biddick, K. Shifting Scales on Common Ground: Developing Personal Expressions and Public Opinions. In International Journal of Continuing Engineering Education and Life Long Learning (2004) 484-505.

[8] Blinkenlights art installation http://www.blinkenlights.de

[9] Obara, H., Tosa, N., and Minoh, M. 2007. Hitch haiku: an interactive generation system of haiku. In Proceedings of the international Conference on Advances in Computer Entertainment Technology (Salzburg, Austria, June 13 - 15, 2007). ACE '07, vol. 203. ACM Press, New York, NY, 248-249.

[10] N. Tosa, A. R. Torroella, B. Ellis, S. Matsuoka, and R. Nakatsu. Computing inspiration: i.plot. In SIGGRAPH '05: ACM SIGGRAPH 2005 Emerging technologies, page 3, New York, NY, USA, 2005. ACM Press.

[11] Charles O. Hartman. Virtual Muse: Experiments in Computer Poetry. Wesleyan University Press, 1996.

[12] Joseph Weizenbaum. Eliza – a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1):36-45, 1966.

[13] Luigi Caputo, Robby Garner and Paco Xander Nathan. FRED, Milton and Barry: the evolution of intelligent agents for the Web. In Advances in intelligent systems (1997) 400-407

[14] Hisar Manurung, Graeme Ritchie and Henry Thompson. Towards a computational model of poetry generation. In Informatics Research Report, Edinburgh, 2000.

[15] Ted Pedersen, Siddharth Patwardhan, Santanjeev Banerjee. Using Measures of Sematic Relatedness For Word Sense Disambiguation.

[16] C. Fellbaum, editor. WordNet: An electronic lexical database. s.l. : MIT Press, 1998.

[17] Siddharth Patwardhan, Ted Pedersen. Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts.